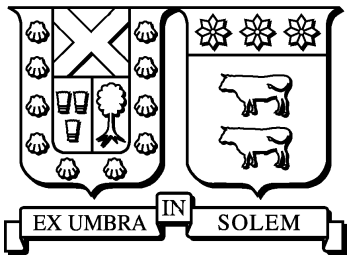
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE INFORMÁTICA

SANTIAGO – CHILE



"PERFILAMIENTO DE ESTUDIANTES Y ESCUELAS A TRAVÉS DE MODELOS DE APRENDIZAJE AUTOMÁTICO PARA LA ELECCIÓN DE ESCUELAS EN CHILE"

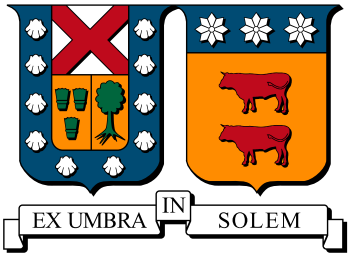
CARLOS ANDRÉS VARGAS POBLETE

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INFORMÁTICO

PROFESOR GUÍA: LIOUBOV DOMBROVSKAIA

JUNIO 2018

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
SANTIAGO – CHILE



"PERFILAMIENTO DE ESTUDIANTES Y ESCUELAS A TRAVÉS DE MODELOS DE APRENDIZAJE AUTOMÁTICO PARA LA ELECCIÓN DE ESCUELAS EN CHILE"

CARLOS ANDRÉS VARGAS POBLETE

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INFORMÁTICO

PROFESOR GUÍA: LIOUBOV DOMBROVSKAIA
PROFESOR CORREFERENTE: PATRICIO RODRÍGUEZ

JUNIO 2018

MATERIAL DE REFERENCIA, SU USO NO INVOLUCRA RESPONSABILIDAD DEL AUTOR O DE LA INSTITUCIÓN

Resumen

Las políticas educacionales buscan asignar recursos públicos al desarrollo de establecimientos educacionales para lograr mejorar el aprendizaje de los niños. Dado que los recursos son escasos, se debe focalizar su asignación en los grupos más necesitados.

Actualmente, los colegios se clasifican por su dependencia administrativa y los alumnos por su nivel socioeconómico. Esta separación unidimensional no representa adecuadamente la realidad.

Los algoritmos de agrupamiento pueden ayudar a definir los programas educativos y sus beneficiarios. Se usaron algoritmos de X-Means modificados, logrando determinar cuatro clústers óptimos para cada entidad. Para los establecimientos los atributos más importantes son la dependencia y el nivel de copago que presentan. En el caso de los estudiantes los atributos más destacados dependerán del enfoque que se le dé, estos pueden ser por un lado el género y el nivel socioeconómico o el tipo de colegio al que asisten y el nivel de copago.

Palabras Clave: Análisis de datos, Aprendizaje No Supervisado, X-Means.

Abstract

Educational policies seek to assign public resources to educational establishments for development, in order to improve children's learning. Given that resources are limited, the designation should be focused on the most vulnerable groups.

Currently, schools are classified by how they are supported financially and students by their socioeconomic level. This one- dimensional segregation does not represent reality accurately.

Clustering algorithms may help define educational programs and their beneficiaries. Using X-Means modified algorithms; four optimal clusters were identified for each entity. For educational establishment the most relevant attributes were financial capacity of the school, private or public, and tuition expenses. For students attributes depend on the perspective, being the most relevant, gender and socioeconomic level or type of school and tuition.

Keywords: Data Analysis, Unsupervised Learning, X-Means.

Índice de Contenidos

Resumen III

Abstract IV

Índice de Contenidos V

Lista de Tablas VII

Lista de Figuras VIII

Glosario X

Introducción 1

1. Definición del Problema 3

2. Estado del Arte 5

2.1. Machine Learning 8

2.2. Clustering 9

2.2.1. K-Means 11

2.2.2. X-Means 12

2.3. CRISP-DM 15

3. Propuesta de Solución **18**

3.1. Metodología de trabajo 18

4. Resultados y Análisis **24**

4.1. Resultados obtenidos 24

4.2. Análisis de los resultados 29

4.2.1. Análisis cualitativo 29

4.2.2. Análisis geográfico 35

Conclusiones **54**

Bibliografía **57**

Índice de tablas

3.1. Atributos que componen la base de datos de los establecimientos. 19

3.2. Atributos que componen la base de datos de las matrículas. 20

3.3. Atributos de relación establecimiento-matrícula para la base de datos de los establecimientos. 21

3.4. Atributos de relación establecimiento-matrícula para la base de datos de las matrículas. 22

4.1. Comparación de clústers de establecimientos por atributo. 25

4.2. Clústers de establecimientos según su dependencia. 25

4.3. Comparación de clústers de establecimientos por atributo, incluyendo los de relación establecimiento-matrícula. 26

4.4. Clústers de establecimientos (tabla 4.3) según su dependencia. 27

4.5. Comparación de clústers de matrículas por atributo. 27

4.6. Clúster de matrículas con atributos de estabilidad inferior al 80 %. 28

4.7. Comparativa de clústers de matrículas por atributo, incluyendo los de relación establecimiento-matrícula. 28

4.8. Clúster de matrículas con atributos de estabilidad inferior al 80 % y atributos de relación. 29

Índice de figuras

2.1. Gráfico de 8 objetos [13]. 10

2.2. Iteración X-Means. 14

2.3. Diagrama de proceso que muestra la relación entre las diferentes fases de CRISP-DM. 16

4.1. Promedios de atributos normalizados de establecimientos de la Región Metropolitana. 31

4.2. Promedios de atributos normalizados de establecimientos con atributos relacionales de la Región Metropolitana. 31

4.3. Promedio de atributos normalizados de matrículas de la Región Metropolitana. 33

4.4. Promedio de atributos normalizados de matrículas con atributos relacionales de la Región Metropolitana. 33

4.5. Mapas de clústers de establecimientos sobre mapa GSE de la Región Metropolitana. 36

4.6. Mapas de clústers de establecimientos (con atributos relacionales) sobre mapa GSE de la Región Metropolitana. 37

4.7. Matrículas en colegios del clúster Municipal Mat. Total Media Baja. 38

4.8. Matrículas en colegios del clúster Subvencionado Mat. Total Media Alta. . 39

4.9. Matrículas en colegios del clúster Pagado Mat. Total Baja. 40

4.10. Matrículas en colegios del clúster Pagado Mat. Total Alta. 40

4.11. Mapas de calor de matrículas en clústers de establecimientos sobre mapa GSE de la Región Metropolitana. 41

4.12. Matrículas en colegios del clúster Municipal IDE Bajo. 42

4.13. Matrículas en colegios del clúster Subvencionado IDE Medio Bajo. 43

4.14. Matrículas en colegios del clúster Pagado IDE Medio Alto. 44

4.15. Matrículas en colegios del clúster Pagado IDE Alto. 44

4.16. Mapas de calor de matrículas en clústers de establecimientos (con atributos relacionales) sobre mapa GSE de la Región Metropolitana. 45

4.17. Matrículas del clúster Hombres Con Beneficio. 47

4.18. Matrículas del clúster Hombres Sin Beneficio. 47

4.19. Matrículas del clúster Mujeres Sin Beneficio. 48

4.20. Matrículas del clúster Mujeres Con Beneficio. 48

4.21. Mapas de calor de clústers de matrículas sobre mapa GSE de la Región Metropolitana (misma escala de densidad). 49

4.22. Matrículas del clúster Mixto Municipal. 50

4.23. Matrículas del clúster Mixto Subvencionado. 51

4.24. Matrículas del clúster Mixto Pagado IPAE Alto. 52

4.25. Matrículas del clúster Mixto Pagado IPAE Bajo. 52

4.26. Mapas de calor de clústers de matrículas (con atributos relacionales) sobre mapa GSE de la Región Metropolitana (misma escala de densidad). 53

Glosario

BIC: Bayesian Information Criterion (Criterio de información bayesiano).

CIAE: Centro de Investigación Avanzada en Educación.

CRISP-DM: Cross Industry Standard Process for Data Mining.

GSE: Grupo Socioeconómico.

IDE: Índice de Desarrollo de la Educación.

IPAE: Índice Potencial de Atracción Espacial.

MICE: Multiple Imputation by Chained Equations (Imputación múltiple por ecuaciones encadenadas).

SEP: Subvención Escolar Preferencial.

Introducción

En Chile los establecimientos educacionales se encuentran categorizados según su dependencia administrativa en municipal, particular subvencionado, particular pagado y corporación de administración delegada, los cuales en la Región Metropolitana se distribuyen de la siguiente manera 23,8 %, 65,2 %, 10 % y 1 % respectivamente [21]. Por otro lado, los estudiantes matriculados en dichos colegios no poseen una categorización clara, y las clasificaciones más cercanas son por el tipo de colegio al cual asisten o por su nivel socioeconómico. La falta de perfiles de establecimientos y matrículas, basados en múltiples atributos, dificulta la tarea de llevar a cabo políticas públicas enfocadas a ciertos perfiles educacionales y estudiantiles.

Debido a esto es que con este estudio se busca encontrar, a partir de diversos atributos, la estructura que tienen los establecimientos en la Región Metropolitana y los estudiantes que a ellos asisten, de manera de poder realizar una mejor clasificación.

Para esto se utilizó la metodología CRISP-DM, en donde primero se tomaron diferentes fuentes de información, las cuales fueron corregidas y estandarizadas para poder trabajar de manera sencilla con ellas. A continuación, se escogieron datos de interés propios de cada establecimiento o estudiante, para poder estudiar su relevancia y también atributos que las relacionan.

Luego, con el fin de no tener una categorización previa de los datos, se utilizó un algoritmo de aprendizaje no supervisado para poder deducir una clasificación directamente de ellos. Primero, se realizó solo con los atributos propios para así poder seleccionar los más relevantes. Después, el resultado obtenido se compara con los resultados que se obtienen al incluir

los datos de relación, tanto en el caso de los establecimientos como de las matrículas, para analizar si al añadir este tipo de información genera un enriquecimiento de la categorización obtenida anteriormente.

Finalmente, con las geolocalizaciones de establecimientos y estudiantes se generan mapas superpuestos al mapa de grupos socioeconómicos (GSE) de la Región Metropolitana para determinar la relación existente entre los clústers generados y el sector socioeconómico en el cual están situados los colegios y en los lugares que residen los estudiantes.

En el primer capítulo se realiza un acercamiento al problema y los objetivos del estudio. Luego, en el segundo, se expone el estado del arte sobre el problema de elección de colegios y se presenta un breve marco teórico conceptual sobre el aprendizaje automático, el clustering y la metodología CRISP-DM. El tercer capítulo se enfoca en la propuesta de solución planteada para el problema en estudio, describiendo la metodología utilizada para el desarrollo del problema. Finalmente, en el cuarto capítulo, se presentan los resultados obtenidos y los análisis correspondientes, para luego presentar las conclusiones del estudio.

Capítulo 1

Definición del Problema

Uno de los principales problemas que surgen al momento de querer generar políticas públicas, es determinar el público objetivo sobre el cual estas se aplicarán, es decir, en que segmento de establecimientos o estudiantes se enfocarán dichas políticas.

En la actualidad, como se señaló anteriormente, los colegios en Chile se clasifican principalmente por la dependencia administrativa, dejando fuera muchos criterios que podrían ser de gran interés. De manera similar, los alumnos que a ellos asisten no cuentan con una clasificación que describa de manera relevante los grupos y sólo se dividen por el tipo de establecimiento al cual asisten o por el nivel de ingreso familiar.

Por esto es necesario identificar los diferentes perfiles de establecimientos y estudiantes presentes en Chile. Para esto se trabajó en conjunto con el Centro de Investigación Avanzada en Educación de la Universidad de Chile, el cual tiene dentro de su misión el dar soporte científico a la discusión y diseño de políticas públicas en el sector educación, para que las políticas nacionales, la gestión educativa local y la docencia en el aula estén basadas en la evidencia que genera la investigación.

El problema al cual se enfrenta el Estado al momento de desarrollar políticas públicas es conocer a cabalidad las características del conjunto o grupo a las cuales estas van dirigidas, incorporando más de una sola característica. Los perfiles, tanto de establecimientos como de estudiantes, los puedan definir de mejor manera un grupo determinado.

El objetivo principal es establecer las relaciones existentes entre perfiles de alumnos y grupos de establecimientos para que a partir de esto se puedan generar políticas públicas de acuerdo a la realidad escolar que se vive en Chile. Además, contar con perfiles establecidos directamente de los datos facilita realizar análisis en diferentes estudios complementarios.

Para llevar a cabo esto se utilizará un algoritmo de clusterización sobre establecimientos y matrículas, que en una primera instancia permitirá determinar los atributos más importantes. En segunda instancia se incorporan atributos que los relacionan y así ver si estos tienen un impacto positivo o negativo en los grupos generados anteriormente.

Frente a la escasez de recursos públicos se debe asegurar que las políticas públicas tengan los efectos esperados. Además, el tener un buen perfil de los grupos significa conocer como estos están formados, lo que apoya la creación de políticas con un impacto positivo. Es decir, el problema afecta a la creación de las políticas y a los efectos que estas causan.

Capítulo 2

Estado del Arte

A la fecha de realización de este estudio no existen trabajos que presenten como objetivo final la clasificación no supervisada de establecimientos educacionales y los alumnos matriculados en cada uno de ellos. Debido a esto se investigó sobre trabajos anteriores en el ámbito de la educación, en específico la elección de colegios, para así tener una guía de que atributos son relevantes de considerar.

En el 2002 Sapelli y Torche [20] estudiaron los diferentes determinantes que inciden en la elección del tipo de colegio que toman en cuenta los padres al momento de matricular a sus hijos en un determinado establecimiento educacional. En este suponen que una mejor educación de los hijos proveerá a los padres una probabilidad mayor de apoyo cuando estén en la vejez, por lo cual utilizan un modelo en donde se postula una función de utilidad para los padres. Dicha función depende del capital humano inicial de los hijos, el cual puede incrementarse con la educación, y del nivel de consumo presente. Para esto utilizaron diferentes fuentes de información, siendo las más relevantes la encuesta de caracterización socioeconómica (CASEN) de 1996 [7] y los resultados del SIMCE, en donde solo consideraron los datos referentes a la elección de establecimientos de enseñanza básica (niños entre 7 y 14 años), en donde el nivel de cobertura es cercano al 100 %[1] y se descarta la opción de no elegir un colegio. Los factores más determinantes son el nivel de ingreso, la educación de los padres, la recepción de subsidios y la calidad del colegio. Los subsidios por colegios

[1] 98,2 % de cobertura educacional para el nivel de enseñanza básica en el año 1996. Fuente: CASEN 1996.

y no por alumno, es decir no son portables, dificultan el acceso a los colegios con menores subsidios para las familias de menores recursos. Otro punto importante que destacan es la alta sensibilidad que los padres demuestran respecto a la calidad de los colegios, aún sin conocer los resultados SIMCE, actúan de tal forma que hace pensar que los conocieran.

En el año 2009 Gallego y Hernando [9] buscando resolver la interrogante de cómo los padres escogen el colegio para sus hijos mejorando al modelo desarrollado anteriormente. Para el estudio se consideraron diferentes atributos, los cuales se pueden agrupar en dos grandes categorías: características del alumno y características del colegio. En ambos casos los datos son obtenidos del SIMCE del 2012 o calculados por los autores a partir de dichos datos para un universo de 70.000 alumnos de cuarto básico que asisten a 1.200 colegios. Los dos atributos que afectan más la elección de un colegio, son el resultado del establecimiento en las pruebas y la distancia entre el hogar y el colegio, en donde el primero se repite respecto al estudio [20].

Dos años después, en el 2011, Daniel Gómez en conjunto con R. Chumacero y R. Paredes [5] realizan un estudio similar a los ya presentados, en donde estudian diversos factores que consideran los padres al escoger un determinado colegio. Dichos factores se pueden clasificar en características particulares de cada niño, las propias de cada establecimiento y las que asocian al niño con la escuela, como la distancia entre el hogar y el colegio. Para llevar a cabo esto establecieron una función similar a la presentada en [20], donde se mide la utilidad de que un niño asista a un determinado colegio y que depende de los tres grupos de factores mencionados. Al igual que en trabajos anteriores fueron usados datos de la encuesta CASEN y del SIMCE, ambos correspondientes al año 2003. Mediante los estudios realizados llegaron a la conclusión de que de los factores analizados, la localización, el precio, la calidad y la potencial competencia de los establecimientos son determinantes al momento de realizar la elección, pero los más valorados por los padres son la calidad y la distancia.

Al año siguiente Gómez, Chumacero y Paredes [10], buscan determinar si el conocimiento de resultados de pruebas específicas (SIMCE) determina de manera importante la selección que realizan los padres sobre el colegio donde matricular a sus hijos. Para esto realizaron un estudio comparativo, tomando como base el estudio anterior y comparándolo con datos de 1996 (primer año donde se hicieron públicos los resultados del SIMCE, por lo cual no

influyen en la elección de colegios de ese año). Del estudio se obtuvo que aún sin conocer los resultados los padres actúan como si los conocieran escogiendo escuelas de mayor calidad, tal como se obtuvo en [20]. Además, cuando los resultados de las pruebas se hicieron públicos, este pasó a ser un factor aún más determinante al momento de tomar una decisión.

Finalmente, uno de los trabajos más recientes en torno a la selección de colegios fue realizado por Canales, Bellei y Orellana [3], donde a diferencia de los trabajos anteriormente señalados, este se enfoca en un sector social específico para determinar y comprender el sentido que tiene para los padres de clase media el elegir un colegio privado. Para este estudio utilizaron dos técnicas complementarias: grupo de discusión y entrevista focalizada, donde la primera apunta a conocer cuál es el valor o significado colectivo de la decisión y la segunda permite conocer como el sujeto entiende la decisión que esta tomando. Los resultados obtenidos son de un carácter preocupante, ya que la selección de colegios esta guiada por el interés del sector medio de distanciarse y diferenciarse de los más pobres, siendo esto una decisión netamente clasista. Además esta preocupa del lado de la educación, debido a que al parecer ni familias ni escuelas parecen orientadas a mejorar el nivel de educación.

Sumado a lo anterior, en el ámbito internacional Abdulkadiroglu y Sonmez [1] en el 2003 propusieron 2 mecanismos de asignación de estudiantes útiles para abordar el problema de elección de colegios, estos son *Gale-Shapley Student Optimal Stable Mechanism* y *Top Trading Cycles Mechanism* (TTC). Ambos mecanismos se basan en las preferencias de los estudiantes y las prioridades que estos tienen en cada colegio. El primero de ellos, basado en lo propuesto en [8] para el problema de admisión a la universidad, asigna a los estudiantes en los establecimientos de su primera preferencia según la prioridad del colegio. En el caso de existir alumnos restantes se procede a revisar su siguiente preferencia y cada colegio vuelve a evaluar por prioridad a los nuevos alumnos y a los que ya estaban asignados, repitiendo dicho procedimiento hasta que todos hayan sido asignados. Con este mecanismo se logra eliminar la envidia justificada[2]. Por otro lado, el segundo mecanismo en términos generales va asignando a los estudiantes por el nivel de prioridad que tienen dentro de las escuelas, donde primero se asignan los de mayor prioridad. Este mecanismo no elimina la envidia justificada, pero si logra reducirla. Este estudio entrega como resultado siete factores en los que

[2] *Justified envy*: Envidia de un estudiante que no fue asignado a la escuela de su preferencia, teniendo una prioridad más alta que la de un alumno que sí fue asignado a dicha escuela.

las decisiones de asignación de estudiantes deberían basarse, las cuales son: balance racial, capacidad de instrucción[3], esfuerzos de replicación, capacidad de estudiantes, prioridad a familias del vecindario, preferencia a hermanos de alumnos del establecimiento y el equilibrio de género.

En el 2006 Chen y Sönmez [4] llevaron a cabo un estudio experimental con los mecanismos de Boston, Gale-Shapley y TTC, donde obtuvieron como resultado que para entornos diseñados[4] el mecanismo de Gale-Shapley es más eficiente que TTC, el cual a su vez es más eficiente que el de Boston. Por otro lado, en entornos aleatorios[5] Boston y Gale-Shapley tienen el mismo nivel de eficiencia, el cual es mayor que el de TTC.

Sumado a los trabajos anteriormente señalados es importante conocer algunos conceptos que son clave para estudios de estás características.

2.1. Machine Learning

El *Machine Learning* [11] o aprendizaje automático es una subrama de la inteligencia artificial que permite a un sistema aprender a través de los datos y no mediante una programación explícita. Este usa diferentes algoritmos que de manera iterativa aprenden de los datos, logrando mejorar la descripción de los datos y la predicción de resultados. Dichos algoritmos operan mediante la construcción de un modelo basado en conjuntos de datos de entrenamiento, que al ir iterando va consiguiendo modelos más precisos. Finalmente, se obtiene un modelo de aprendizaje automático que es la salida generada al entrenar un algoritmo de aprendizaje con datos.

En otras palabras es una ciencia que permite el estudio del comportamiento o patrones presentes en diversos tipos de datos, para poder automatizar diversos procesos. Además, esto

[3]Recursos que la escuela posee para apoyar la enseñanza. La capacidad o poder para utilizar estos recursos de la manera más efectiva.

[4]Entorno "realista" en donde existe una correlación de las preferencias de los estudiantes con la proximidad y la calidad de la escuela.

[5]Entorno donde las preferencias no están correlacionadas.

implica un aprendizaje continuo que va mejorando con cada iteración haciéndolo más inteligente y capaz de resolver diferentes problemas en base a lo que va aprendiendo.

Su utilización en diversos ámbitos tiene variadas ventajas, esta permite mejorar la gestión organizacional, facilitar la toma de diferentes decisiones, automatizar y acelerar procesos, entre muchas otras. Pero como es de esperarse también presenta desventajas, las cuales vienen muy de la mano con las decisiones humanas, ya que estas decisiones afectan la resolución que toma el algoritmo en las tareas que se le asignan. Una mala decisión humana puede influir en malos resultados del algoritmo y un mal desempeño en las tareas asignadas.

Este aprendizaje se divide en dos: supervisado y no supervisado.

El aprendizaje supervisado consiste en intentar deducir a partir de datos de entrenamiento o ejemplo una función que clasifique datos sin una clasificación previa. En este caso los datos están compuestos por dos partes, por un lado están los diferentes atributos, ya sean numéricos o categóricos, y por otro lado una etiqueta que clasifica el dato. Entonces, lo que se hace es determinar mediante los diferentes atributos el valor de la etiqueta, para luego poder predecir la clasificación de datos no categorizados. Un ejemplo de algoritmo de aprendizaje no supervisado es el de K vecinos mas cercanos [2].

El aprendizaje no supervisado, a diferencia del supervisado no posee un conocimiento a priori de una clasificación de los datos. Por lo tanto, un modelo se ajusta al número de observaciones que contiene un data set. En este tipo de aprendizaje solo se cuenta con diferentes atributos, sobre los cuales se buscan semejanzas para poder clasificarlos y crear agrupaciones o clústers. Algunos ejemplos de algoritmos de aprendizaje no supervisado son K-Means [14] y la extensión propuesta en [18], X-Means.

2.2. Clustering

El agrupamiento o *clustering* [13] es un procedimiento que consiste en buscar grupos en un conjunto de datos, estos se denominan clústers y encontrarlos es el objetivo principal del análisis de clústers. Lo que se busca con esto es básicamente formar grupos que contengan

objetos similares entre sí y que sean lo más diferente posible con los objetos de otros grupos.

Esta tarea es una actividad importante del proceso de aprendizaje del ser humano, la cual comienza desde muy pequeños. Tareas como la distinción entre perros y gatos u hombres y mujeres, entre muchos otros, son claros ejemplos de esto. La clasificación siempre a jugado un papel muy importante dentro de la ciencia y es por esto que existen innumerables aplicaciones, tales como la clasificación de compuesto en química, la identificación de diversas enfermedades en la medicina, la clasificación de estrellas en la astronomía, etc.

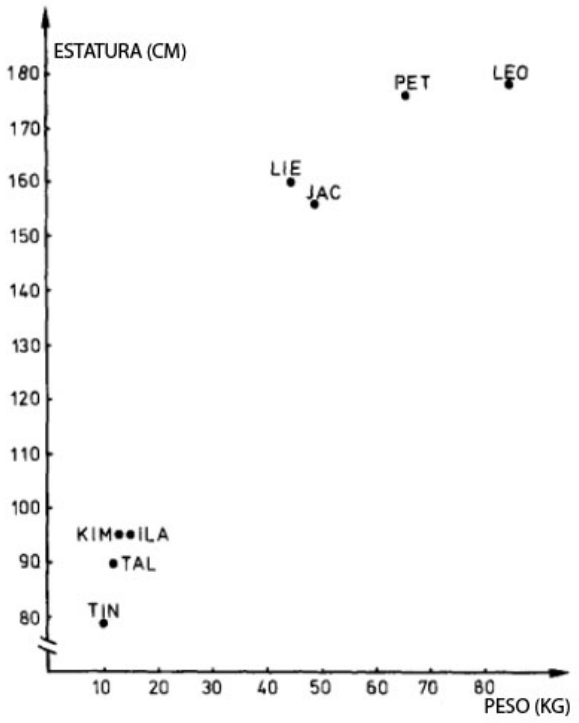


Figura 2.1: Gráfico de 8 objetos [13].

En el pasado los agrupamientos se realizaban generalmente de manera subjetiva, basándose en el juicio y la percepción del observador. Esto es posible de realizar para casos simples con 2 o 3 dimensiones, como el de la figura 2.1, donde se pueden reconocer a simple vista 2 grupos {ILA, KIM, TIN, TAL} y {JAC, LEO, LIE, PET}. Sin embargo, en la vida cotidiana es necesario clasificar objetos con una mayor dimensionalidad y realizar este procedimiento de manera más objetiva. Es por esto que se ha desarrollado el procedimiento automático de clasificación.

Con los años se han desarrollado múltiples algoritmos de clasificación, debido a que no existe una definición general de clúster y de hecho existen muchos tipos: clústers esféricos, clústers lineales, entre otros. Además, cada una de las distintas aplicaciones utilizan diferentes tipos de datos como variables discretas, variables continuas, similitudes y diferencias.

A continuación se presentan los algoritmos de *clustering* utilizados.

2.2.1. K-Means

Este es un método de agrupamiento que tiene por finalidad clasificar un grupo de datos en k grupos, basándose en la distancia que presenta cada dato respecto a un centroide. Este algoritmo fue propuesto por Lloyd en 1957, pero no fue publicado hasta el año 1982 [14].

Para realizar el agrupamiento el algoritmo realiza iteraciones, en donde lleva a cabo un seguimiento de los centroides de los grupos. Antes de la primera iteración se inicializan los centroides con valores aleatorios. En cada iteración se realizan las siguientes tareas. Primero, para cada punto de los datos se encuentra el centroide más cercano y se asocia a este. Después se reestiman las ubicaciones de los centroides con el centro de masa de sus puntos asociados. Luego se repite el mismo proceso de asignar cada punto al centroide más cercano y reestimar la ubicación de los centroides hasta que estos permanezcan fijos en una iteración o se alcance un número determinado de iteraciones.

En el algoritmo 1 se muestra el pseudocódigo del algoritmo de agrupamiento descrito.

Las principales ventajas que presenta este algoritmo son: facilidad para ser implementado, simple y general. Por otro lado, sus desventajas son: se debe especificar un k, muy sensible a la inicialización de los centroides y su baja escalabilidad. Una mala elección del número de clústers puede repercutir en la obtención de malos resultados, al igual que la inicialización aleatoria de los centroides. Además, mientras más grande es el número de clústers, mayor es la probabilidad de incurrir en mínimos locales.

Algoritmo 1 K-Means [17]

Require: $K \geq 1$

1: **for all** i ∈ [1..K] **do**
2: $c_i \leftarrow$ ObtenerPuntoAleatorio()
3: **repeat**
4: Set all C_i equal to $\emptyset$
5: **for all** item ∈ S **do**
6: $j \leftarrow \text{argmin}_i(\|c_i - \text{item}\|)$
7: $C_j = C_j \cup \{\text{item}\}$
8: **for all** i ∈ [1..K] **do**
9: $\text{old\_}c_i = c_i$
10: $c_i = \frac{1}{|C_i|} \sum_{j \in C_i} j$
11: **until** $\forall_i, \text{old\_}c_i = c_i$
12: **return** all C_i

2.2.2. X-Means

X-Means es un algoritmo de agrupamiento, extendido de K-Means, propuesto por Pelleg y Moore en [18], el cual busca mejorar la baja escalabilidad computacional, uso de un número determinado de clústers y sensibilidad a mínimos locales.

El principal problema que viene a solucionar este método es el de ingresar con anticipación el número deseado de clústers. A diferencia de K-Means, X-Means recibe un límite inferior y uno superior, dentro del cual el algoritmo es capaz de determinar cual es el número de centroides correcto basándose en una heurística.

El algoritmo 2 generado por Montresor y Guerrieri [17] muestra el funcionamiento de X-Means, el cual está basado en un K-Means reiterativo con $K = 2$. Lo que realiza este método es dividir en dos el data set inicial, para luego ir dividiendo en dos cada clúster que se va generando y detenerse cuando el número de clústers es mayor al límite superior.

En palabras sencillas el algoritmo realiza las siguientes operaciones:

1. Ejecuta K-Means ($K = 2$) en el conjunto completo de datos, tomando dos centroides a partir de un vector aleatorio que pasa por el centro de masa del conjunto original y a una distancia proporcional al tamaño de la región total.

2. Si los clústers ”hijos” tienen un desempeño mejor según el criterio de información bayesiano (BIC) que el clúster original, estos se conservan y lo reemplazan.

3. Si no existe una mejor representación del clúster original escoge una fracción constante de los clústers y los reemplaza por sus dos ”hijos”.

4. El algoritmo se detiene cuando el número de clústers es mayor al límite superior entregado al algoritmo.

Algoritmo 2 X-Means (simplificado) [17]

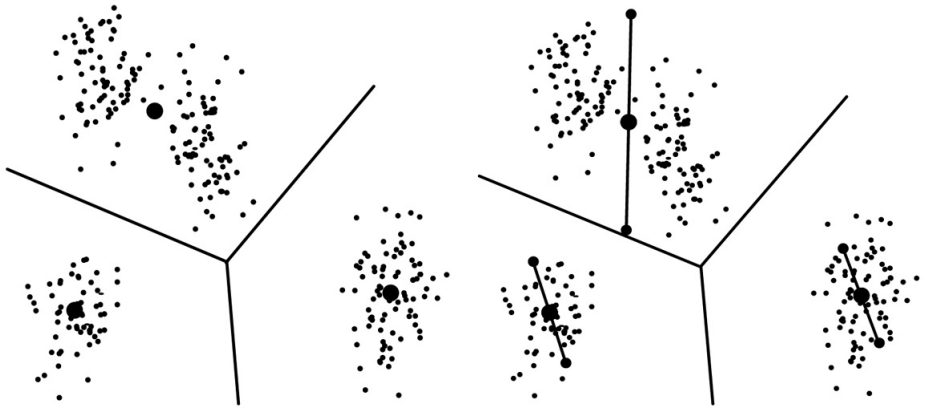
Require: Set de datos S, número máximo de cluster MAX

Require: Función 2Means(S) retorna 2 clústers

```
1:  Clustering ← 2Means(S)
2:  Mejor_Puntuacion ← −∞
3:  while |Clustering|< MAX do
4:      Nuevo_Clustering ← {}
5:      for all Cl ∈ Clustering do
6:          Cl2 ← 2Means(Cl)
7:          if Medida(Cl) > Medida(Cl2) then
8:              Nuevo_Clustering ← Nuevo_Clustering ∪ {Cl}
9:          else
10:             Nuevo_Clustering ← Nuevo_Clustering ∪ Cl2
11:     Clustering ← Nuevo_Clustering
12:     if Measure(Clustering) > Mejor_Puntuacion then
13:         Mejor_Puntuacion ← Medida(Clustering)
14:         Mejor_Clustering ← Clustering
15: return Mejor_Clustering
```

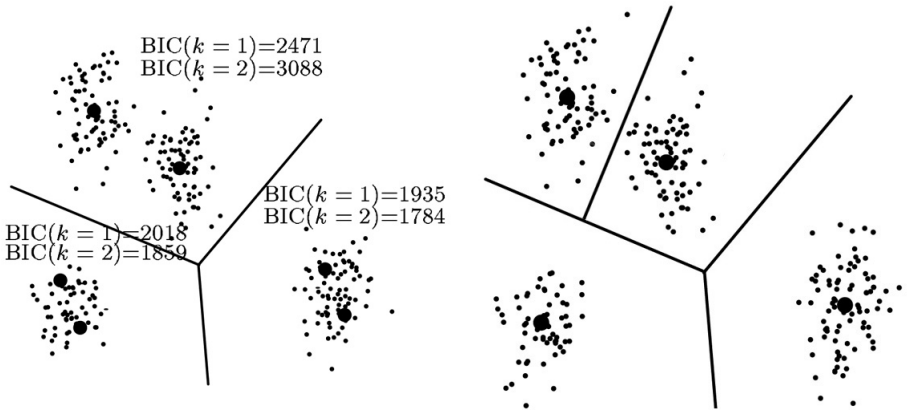
Lo anteriormente descrito se puede apreciar de forma gráfica en la figura 2.2, donde se encuentra una representación para una iteración del algoritmo, con un conjunto inicial de 3

clústers.



(a) El resultado de ejecutar K-Means con 3 centroides.

(b) Cada centroide original se divide en dos hijos.



(c) El resultado después de que todos los 2-Means paralelos hayan terminado.

(d) Los centroides sobrevivientes después de todas las pruebas de puntuación.

Figura 2.2: Iteración X-Means.

En la figura 2.2a se muestra una solución de K-Means con 3 centroides y los límites de cada región. En la 2.2b se dividen los centroides en dos nodos hijos, trazando un vector aleatorio a una distancia proporcional al tamaño de la región. Luego de esto se ejecuta un K-Means con $k = 2$ de manera local, esto debido a que los hijos solo se distribuyen los puntos de los padres y no se relacionan con el resto. El resultado al finalizar los K-Means se muestra en la figura 2.2c, donde cada hijo tiene sus puntos asociados.

Una vez finalizado lo anterior es momento de evaluar si el padre o los hijos representan mejor la distribución de cada región. Según este criterio se eliminará el padre o su descendencia. Si la representación original de una región es la adecuada, entonces sobrevivirá el padre y se matará a sus hijos. Por otro lado, las regiones que no estén bien representadas por los centroides recibirán mayor atención, aumentando en ellas el número de centroides. En la figura 2.2d se muestra el resultado final al llevar a cabo la prueba de puntuación (BIC) en los 3 pares de hijos.

Una de las principales ventajas que posee este algoritmo radica en que es más escalable debido a que con cada iteración se reduce más el número de datos en los cuales K-Means se debe ejecutar, lo que hace que sea más fácil utilizarlo en conjuntos de datos de mayor tamaño. Otra ventaja es que se sabe que con K-Means de pocos clústers es menos probable incurrir en mínimos locales en comparación a uno realizado con muchos clústers. Por lo tanto, el hecho de que X-Means utilice un K-Means con $K = 2$ favorece a que no se atasque en mínimos locales.

Además este método permite realizar una visualización del tipo árbol, la que crea una estructura jerárquica de los clústers.

El porqué se utilizará este algoritmo es debido a su sencillez y facilidad de implementación y ejecución. Además, otro punto importante por el cual se optó por este algoritmo es porque el problema que se busca resolver cuenta con una gran cantidad de datos, lo que no es una dificultad debido a la escalabilidad que este posee gracias a que con cada iteración el conjunto de puntos sobre el cual se aplica K-Means ($k = 2$) se ve reducido. Por último, pero no menos importante, se escogió este algoritmo debido a que no necesita conocer de antemano un número de clústers, sino que es calculado automáticamente.

2.3. CRISP-DM

CRISP-DM [22] (*Cross Industry Standard Process for Data Mining*) es una de las metodologías que más se utiliza para el desarrollo de proyectos de minería de datos. Esta proporciona una visión general del ciclo de vida de un proyecto de minería de datos. Además contiene

las fases de un proyecto, sus respectivas tareas y sus resultados.

Este ciclo esta compuesto por 6 fases que se aprecian en la figura 2.3. Dentro del diagrama las flechas sólo indican las dependencias más importantes y frecuentes entre las fases, en donde la secuencia no es estricta. En la figura 2.3 el círculo exterior representa la naturaleza cíclica de la minería de datos. Este proceso no termina una vez se implemente la solución, debido a que todo lo aprendido durante el proyecto puede desencadenar nuevas preguntas de negocio, en donde dichos procesos se beneficiarán de experiencias anteriores.

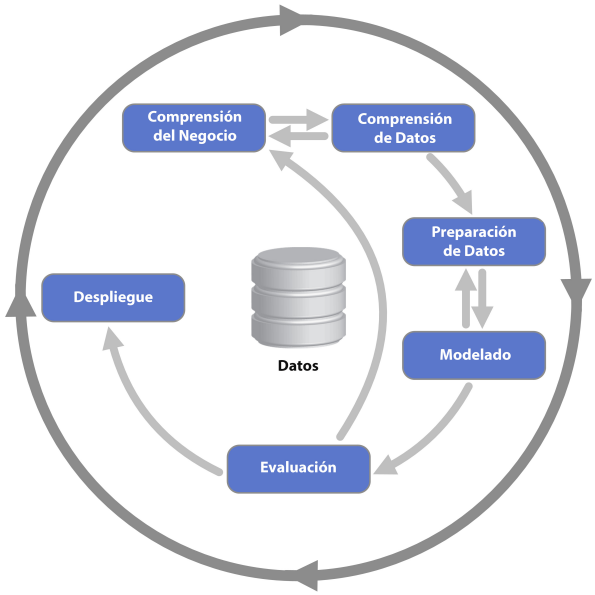


Figura 2.3: Diagrama de proceso que muestra la relación entre las diferentes fases de CRISP-DM.

1. Comprensión del negocio: Esta fase se centra en comprender los objetivos y requisitos del proyecto desde una perspectiva comercial, para luego convertirlo en una definición del problema de minería de datos. Además, crear un plan preliminar diseñado para alcanzar los objetivos.

2. Comprensión de datos: Esta fase comienza con una recopilación inicial de datos y actividades para familiarizarse con los datos, identificar problemas de calidad de datos, descubrir primeras ideas sobre los datos o detectar subconjuntos interesantes para formar una hipótesis de información oculta.

3. Preparación de datos: Esta cubre todas las actividades para construir el conjunto de datos final a partir de los datos brutos iniciales. Dichas tareas incluyen la selección y transformación de tablas, registros y atributos, construcción de nuevos atributos y limpieza de datos para las herramientas de modelado. Las tareas de preparación de datos puede que se realicen varias veces y sin un orden específico.

4. Modelado: Fase de selección y aplicación de diversas técnicas de modelado, además de la calibración de sus parámetros para obtener óptimos resultados. Existen varias técnicas para el mismo tipo de problemas de minería de datos. Algunas de ellas requieren formatos de datos específicos.

5. Evaluación: En esta etapa se ha creado uno o varios modelos que parecen ser de alta calidad. Antes de realizar el despliegue final es importante evaluar más a fondo el modelo y revisar los pasos ejecutados para construirlo.

6. Despliegue: En esta fase se deberá definir como implementar los resultados. Dicha implementación dependerá de los requisitos, esta puede ser tan simple como generar un informe o tan compleja como implementar un proceso repetible de minería de datos.

Capítulo 3

Propuesta de Solución

Este trabajo presenta un estudio tanto de establecimientos educacionales como de sus alumnos, para poder determinar la estructura que estos presentan y poder generar un modelo de selección de colegios basado en los grupos que se generen a partir de los conjuntos de datos iniciales.

3.1. Metodología de trabajo

Con la finalidad de llevar a cabo los objetivos planteados para este estudio se optó por utilizar la metodología CRISP-DM, la cual es la más usada por los expertos en minería de datos.

El trabajo se desarrolla de la siguiente forma:

- **Comprensión de datos:** esta etapa consiste básicamente en la recolección de datos desde diferentes fuentes. En este caso fueron utilizadas las bases de datos de establecimientos y matrículas de la Región Metropolitana facilitadas por el CIAE. Además, se complementó la base de datos de los colegios con la información pública disponible en la página Web del Ministerio de Educación de Chile (MIME [6]), los cuales fueron obtenidos mediante la técnica de *Web scraping* [16].

- **Preparación de datos:** en esta fase, a partir de los datos obtenidos anteriormente, se seleccionaron los atributos más relevantes para el estudio y se realizó un proceso de limpieza y estandarización de los datos con R. Además, se agregaron algunos atributos derivados o calculados de los datos extraídos originalmente.

  Para cada atributo se calculó su nivel de estabilidad, la cual mide cuán estable o constante es el atributo. La estabilidad se calcula mediante la división del número de filas no nulas del valor más frecuente y el número total de filas de datos no nulos.

  Lo anteriormente descrito se resume en las tablas 3.1 y 3.2, las cuales presentan los atributos correspondientes a los establecimientos y matrículas, respectivamente, ordenados de menor a mayor según su estabilidad.

Tabla 3.1: Atributos que componen la base de datos de los establecimientos.

Atributo	**Descripción**	**Estabilidad**	**Fuente**
mat_total	Matrícula total de alumnos.	0,44 %	MIME
prom_alu_cur	Promedio de alumnos por curso.	5,22 %	MIME
nivel_deportivo	Nivel deportivo del establecimiento.	48,36 %	MIME
nivel_ensenanza	Nivel de enseñanza que ofrece el establecimiento.	48,4 %	Derivado de datos CIAE.
cod_depe	Código de dependencia del establecimiento.	53,82 %	CIAE
enf_valorico	Enfoque valórico.	54,79 %	MIME
pago_men	Nivel de pago de mensualidad.	60,64 %	MIME
req_entrevista	Requisitos de entrevista para postular.	62,81 %	MIME
becas_disp	Becas disponibles en el establecimiento.	64,07 %	MIME
convenio_sep	Posee convenio de subvención escolar preferencial (SEP).	71,91 %	MIME
pago_mat	Nivel de pago de matrícula.	77,66 %	MIME
enf_academico	Enfoque académico.	84,62 %	MIME
area_metropolitana_rbd	Pertenencia al área metropolitana.	85,74 %	CIAE
continúa . . .			

Tabla 3.1: Atributos que componen la base de datos de los establecimientos (continuación).

apoyo_tutorias	Ofrece ayuda a los alumnos mediante tutorías.	86,65 %	MIME
req_papeles	Requisitos de papeles para postular.	87,57 %	MIME
req_pago	Requisitos de pago para postular.	89,12 %	MIME
apoyo_especialistas	Ofrece ayuda a los alumnos mediante especialistas.	90,18 %	MIME
gen_rbd	Género del establecimiento.	94,54 %	CIAE
req_otros	Requisitos de cualquier tipo que no clasifique en las otras categorías.	95,02 %	MIME
enf_laboral	Enfoque laboral.	95,21 %	MIME
req_prueba	Requisitos de prueba para postular.	95,94 %	MIME
enf_otros	Enfoque de otro tipo que no clasifique en las otras categorías.	96,32 %	MIME
apoyo_otros	Ofrece ayuda a los alumnos de cualquier otra forma que no clasifique en las otras categorías.	98,79 %	MIME

Tabla 3.2: Atributos que componen la base de datos de las matrículas.

Atributo	**Descripción**	**Estabilidad**	**Fuente**
IPAE	Indicador de barrio crítico.	0,6 %	CIAE
edad_alu	Edad del alumno.	8,79 %	CIAE
gen_alu	Género del alumno.	9,04 %	CIAE
criterio_sep	Criterio por el cual se considera prioritario.	9,23 %	CIAE
beneficiario_sep	Indicador del alumno beneficiario de la SEP.	50,73 %	CIAE
gse_decil_nac	Grupo Socioeconómico del alumno.	64,04 %	CIAE
continúa . . .			

Tabla 3.2: Atributos que componen la base de datos de las matrículas (continuación).

sobre_edad	Diferencia entra la edad actual y la esperada para el nivel.	68,95 %	Calculado con datos CIAE.
area_metropolitana_alu	Pertenencia al área metropolitana.	90,48 %	CIAE
cod_sec	Código del sector económico.	94,81 %	CIAE
cod_espe	Código de especialidad.	94,81 %	CIAE
cod_rama	Código de rama.	94,81 %	CIAE
grado_sep	Corresponde a un nivel SEP.	100 %	CIAE

Además de los atributos propios de cada establecimiento y matrículas se añadieron variables que relacionan a ambas entidades, las variables de relación establecimiento-matrícula. En el caso de los establecimientos se agregó la sobre edad promedio, el GSE que predomina en el establecimiento, la desviación de la moda GSE, los porcentajes de hombres y mujeres, la distancia máxima a la que viven el 75 % de los alumnos de cada colegio y el índice de desarrollo de la educación (IDE) por rangos. Por otro lado, para las matrículas se añadió la distancia a la que vive el alumno del colegio, la dependencia del colegio y el nivel de copago (matrícula y mensualidad). Estos se pueden ver en las tablas 3.3 y 3.4, respectivamente.

Tabla 3.3: Atributos de relación establecimiento-matrícula para la base de datos de los establecimientos.

Atributo	**Descripción**	**Fuente**
sobre_edad_prom	Sobre edad promedio de los alumnos que estudian en el establecimiento.	Calculado con datos CIAE.
desviacion_moda	Desviación que presenta el grupo GSE con respecto a la moda.	Calculado con datos CIAE.
porc_femenino	Porcentaje de mujeres del establecimiento.	Calculado con datos CIAE.
porc_masculino	Porcentaje de hombres del establecimiento.	Calculado con datos CIAE.
dist_percentil_75	Distancia del percentil 75 de los alumnos que asisten al establecimiento.	Calculado con datos CIAE.
continúa . . .		

Tabla 3.3: Atributos de relación establecimiento-matrícula para la base de datos de los establecimientos (continuación).

IDE_rango	Índice de Desarrollo de la Educación para Todos por rango.	Calculado con datos CIAE.
gse_moda_nac	Grupo Socioeconómico que más se repite en el establecimiento.	Calculado con datos CIAE

Tabla 3.4: Atributos de relación establecimiento-matrícula para la base de datos de las matrículas.

Atributo	**Descripción**	**Fuente**
dist_actual	Distancia del alumno a su establecimiento actual.	CIAE
pago_men	Nivel de pago de mensualidad.	MIME
cod_depe	Dependencia del colegio al que asiste.	CIAE
pago_mat	Nivel de pago de matrícula.	MIME

Adicionalmente se imputaron los campos de datos faltantes mediante el algoritmo MICE (*Multiple Imputation by Chained Equations*) [19] para eliminar los valores nulos.

- **Modelado:** en esta etapa se escoge el algoritmo de agrupamiento X-Means, debido a que uno de los objetivos del estudio es poder encontrar sin una idea previa la estructura que poseen los establecimientos y los alumnos, el cual es ejecutado sobre el software RapidMiner [15]. En el programa se transforman las variables categóricas a numéricas y se realiza una normalización. Además se ajustan los diferentes parámetros del algoritmo y se define que los atributos a utilizar deben tener una estabilidad máxima de 80 %, es decir, se considerarán solo los atributos que tengan una estabilidad inferior al 80 %.

- **Análisis y resultados:** se realiza un análisis de los resultados obtenidos de la fase de modelado, los cuales deben ser capaces de satisfacer los objetivos planteados para el estudio.

- **Conclusiones e implementación:** finalmente se deben realizar las conclusiones respectivas del trabajo efectuado, validar el cumplimiento de los objetivos que fueron planteados para el trabajo y la realización del presente informe.

Capítulo 4

Resultados y Análisis

Una vez realizados los experimentos con el algoritmo de agrupamiento X-Means, se presentan los resultados en diferentes figuras y tablas. Primero se encuentran los resultados conseguidos al ejecutar el algoritmo con los atributos de las tablas 3.1 y 3.2. Después se presentan los resultados obtenidos al incorporar los atributos de relación establecimiento-matrícula (tablas 3.3 y 3.4) en ambas entidades, para luego realizar una comparación de ambos resultados. Finalmente, se realiza un análisis cualitativo de los resultados y un análisis geográfico de estos.

4.1. Resultados obtenidos

De la tabla 4.1 a la 4.4 se presentan los resultados obtenidos al ejecutar X-Means con los datos correspondientes a los establecimientos. En la tabla 4.1 se encuentran los resultados que se obtuvieron al ejecutar el algoritmo de agrupamiento con los atributos propios de los establecimientos que tienen una estabilidad inferior o igual al 80 %. En ella se resume atributo por atributo que tipo de establecimientos se encuentran en cada uno de los 4 clústers generados.

Tabla 4.1: Comparación de clústers de establecimientos por atributo.

	Municipal Mat. Total Media Baja	**Subvencionado Mat. Total Media Alta**	**Pagado Mat. Total Baja**	**Pagado Mat. Total Alta**
Dependencia	Municipal	P. Subvencionado / Corp. Admin. Delegada	P. Pagado	P. Pagado
Nivel impartido	Básica	Básica / Completa	Completa	Completa
Matrícula	Gratuito	Bajo $10.000	Más de $100.000	Más de $100.000
Mensualidad	Gratuito	Bajo $50.000	Más de $100.000	Más de $100.000
Convenio SEP	Si	Si	No	No
Prom. matrículas	480	649	297	1084
Prom. alumnos por curso	27	26	17	26
Prom. de becas	15	60	5	12
Nivel deportivo	Medio	Medio	Bajo	Alto
Req. entrevista	Medio	Medio	Alto	Alto
Enf. valórico	Medio	Bajo	Bajo	Medio

La tabla 4.2 muestra la composición de los 4 clústers obtenidos según la dependencia del establecimiento y el total de colegios que cada uno de los clústers posee.

Tabla 4.2: Clústers de establecimientos según su dependencia.

	Municipal Mat. Total Media Baja	**Subvencionado Mat. Total Media Alta**	**Pagado Mat. Total Baja**	**Pagado Mat. Total Alta**
Municipal	660	1	0	0
P. Subvencionado	0	1112	0	1
P. Pagado	0	0	134	127
Corp. Admin. Delegada	0	33	0	0
TOTAL	660	1146	134	128

En la tabla 4.3 se encuentran los resultados que se obtienen al incluir tanto los atributos propios con estabilidad menor o igual a 80 % y los atributos de relación establecimiento-matrícula de la tabla 3.3. De manera similar a lo realizado al considerar sólo los atributos propios, se resumen las características de los colegios pertenecientes al clúster.

Tabla 4.3: Comparación de clústers de establecimientos por atributo, incluyendo los de relación establecimiento-matrícula.

	Municipal IDE Bajo	**Subvencionado IDE Medio Bajo**	**Pagado IDE Medio Alto**	**Pagado IDE Alto**
Dependencia	Municipal / Corp. Admin. Delegada	P. Subvencionado / Corp. Admin. Delegada	P. Pagado	P. Pagado
Nivel impartido	Básica	Básica / Completa	Completa	Completa
Matrícula	Gratuito	Bajo $10.000	Más de $100.000	Más de $100.000
Mensualidad	Gratuito	Bajo $50.000	Más de $100.000	Más de $100.000
Convenio SEP	Si	Si	No	No
Prom. matrículas	430	692	677	807
Prom. alumnos por curso	25	30	21	26
Prom. de becas	14	62	9	9
Nivel deportivo	Medio	Bajo	Medio	Bajo
Req. entrevista	Medio	Medio	Alto	Alto
Enf. valórico	Medio	Medio	Medio	Alto
IDE	-1 a 0	-0,5 a 0,5	0,5 a 1	1 a 1,5
Sobre edad	0,59	0,34	0,44	0,44
Distancia P. 75	3,4 km	3,6 km	7,8 km	6,1 km
GSE	40	40	40	60
Desviación GSE	5,7	7,5	11,6	13,6
% Hombres	56	49	54	15
% Mujeres	44	51	46	85

La tabla 4.4 muestra la cantidad de colegios por dependencia que tiene cada uno de los 4 clústers obtenidos y el total de establecimientos que lo forman.

Tabla 4.4: Clústers de establecimientos (tabla 4.3) según su dependencia.

	Municipal IDE Bajo	**Subvencionado IDE Medio Bajo**	**Pagado IDE Medio Alto**	**Pagado IDE Alto**
Municipal	640	21	0	0
P. Subvencionado	67	1.042	1	3
P. Pagado	0	0	241	20
Corp. Admin. Delegada	20	13	0	0
TOTAL	727	1.076	242	23

Las siguientes tablas (4.5 a 4.8), y de manera similar a lo descrito anteriormente para los establecimientos, muestran los resultados obtenidos usando los datos de las matrículas. En la tabla 4.5 se muestran los 4 clústers obtenidos como resultado de ejecutar X-Means con los atributos de la tabla 3.2 que presentan un valor inferior o igual al 80 % en su estabilidad.

Tabla 4.5: Comparación de clústers de matrículas por atributo.

	Hombres Con Beneficio	**Hombres Sin Beneficio**	**Mujeres Sin Beneficio**	**Mujeres Con Beneficio**
Promedio edad	11,7	11,66	11,66	11,68
Sobre edad	0,44	0,38	0,29	0,34
Género	Hombres	Hombres	Mujeres	Mujeres
Beneficiario SEP	Si	No	No	Si
GSE	40	80	80	40
IPAE	0,29	0,47	0,47	0,28

La tabla 4.6 presenta el número de alumnos por género que conforman los 4 clústers obtenidos con los atributos propios de las matrículas y el total de alumnos que pertenece a cada uno.

Tabla 4.6: Clúster de matrículas con atributos de estabilidad inferior al 80 %.

	Hombres Con Beneficio	**Hombres Sin Beneficio**	**Mujeres Sin Beneficio**	**Mujeres Con Beneficio**
Hombres	449.815	81.981	0	3
Mujeres	0	0	78.660	437.915
TOTAL	449.815	81.981	78.660	437.918

En la tabla 4.7 se resume por atributo las características de los 4 grupos obtenidos al ejecutar el algoritmo de clustering con los atributos propios con estabilidad menor o igual a 80 % de estabilidad y los atributos relacionales señalados en la tabla 3.4.

Tabla 4.7: Comparativa de clústers de matrículas por atributo, incluyendo los de relación establecimiento-matrícula.

	Mixto Municipal	**Mixto Subvencionado**	**Mixto Pagado IPAE Alto**	**Mixto Pagado IPAE Bajo**
Promedio edad	11,76	11,64	11,59	11,78
Género	Mixto	Mixto	Mixto	Mixto
Beneficiario SEP	Si	Si/No	No	No
GSE	40	50	90	40
IPAE	0,3	0,31	0,56	0,33
Dependencia del establecimiento	Municipal	P. Subvencionado / Corp. Admin. Delegada	P. Pagado	P. Pagado
Distancia	4,7 km	4,2 km	5,2 km	5,5 km
Sobre edad	0,5	0,33	0,4	0,4
Pago matrícula	Gratuito	Bajo $10.000	Sobre $50.000	Sobre $50.000
Pago mensualidad	Gratuito	Bajo $50.000	Más de $100.000	Más de $100.000

Por último, en la tabla 4.8 se presenta la cantidad de hombres y mujeres que tiene cada uno de los 4 clústers que se obtuvieron al ejecutar el algoritmo con los atributos propios y de relación para las matrículas, además se contabiliza el total de alumnos que tienen los clústers.

Tabla 4.8: Clúster de matrículas con atributos de estabilidad inferior al 80 % y atributos de relación.

	Mixto Municipal	**Mixto Subvencionado**	**Mixto Pagado IPAE Alto**	**Mixto Pagado IPAE Bajo**
Hombres	143.945	314.346	18.535	54.973
Mujeres	124.668	319.734	17.734	54.439
TOTAL	268.613	634.080	36.269	109.412

4.2. Análisis de los resultados

4.2.1. Análisis cualitativo

En esta sección se analizan los diferentes resultados obtenidos y presentados anteriormente, comenzando con el análisis para los establecimientos y seguido por el de las matrículas.

A partir de los resultados obtenidos para los establecimientos y de las figuras generadas con estos (figuras 4.1 y 4.2), podemos ver que al realizar una comparación entre ambas ejecuciones del algoritmo (sin y con atributos de relación) existen 2 clústers que sin importar la ejecución, presentan características y cardinalidades similares, la diferencia está en que en la segunda ejecución aumenta el nivel de información del grupo. En el primer caso, el clúster de colegios municipales de matrícula total media baja, generado en la primera ejecución, es similar al clúster de establecimientos municipales de IDE bajo. En ambos se concentran colegios de dependencia municipal, de matrícula y mensualidad gratuita, poseen convenio SEP y en su mayoría el nivel educacional que imparten es enseñanza básica. Estos establecimientos presentan un IDE que va de regular a bajo, lo que indica que su calidad de educación es baja, y un nivel alto de segregación en comparación a los otros clústers, debido a que tiene la desviación del GSE más baja. Además, los alumnos que asisten a estos colegios pertenecen principalmente a un grupo socio económico bajo (percentil 40), el 75 % de ellos se desplaza como máximo 3,4 km para llegar a su lugar de estudio y son el grupo con mayor sobre edad promedio, es decir son los alumnos que más han repetido.

En el segundo caso, el clúster de los subvencionados de matrícula total media alta es similar al de los subvencionados de IDE medio bajo que se obtiene al agregar los atributos relacionales. Estos se componen principalmente por colegios particulares subvencionados de bajo costo en matrículas y mensualidades (no superan los $50.000) que imparten educación básica y completa, principalmente y no poseen convenio SEP. Dichos establecimientos tienen un IDE regular que va entre -0,5 y 0,5, su nivel de sobre edad es el más bajo de todos y es el segundo con menos segregación (desviación GSE baja). Los estudiantes de estos establecimientos, al igual que el grupo anterior, viven a no más de 3,6 km de distancia y en su mayoría pertenecen a un GSE bajo (percentil 40).

Los terceros y cuartos clústers están compuestos por establecimientos particulares pagados con un nivel de copago elevado (superior a $100.000), imparten educación completa (básica y media) y no poseen convenio SEP. Estos se diferencian entre sí, en el caso de la primera ejecución, principalmente por el número de matrículas totales y el número de alumnos por curso. El clúster de colegios pagados de matricula total baja tiene un promedio de 300 alumnos inscritos y cursos de 17 alumnos. El clúster de pagados con matrícula total alta tiene alrededor de 1000 alumnos matriculados y 26 alumnos por curso. En el caso de la segunda ejecución, los atributos que diferencian a ambos grupos son el IDE, la distancia a la que viven los estudiantes y el GSE, en donde el primero (Pagado IDE Medio Alto) tiene una buena calidad de educación (IDE entre 0,5 y 1), con alumnos de familias que en promedio pertenecen al grupo socio económico del percentil 40, pero con muy poca segregación debido a su alto nivel de desviación. Además, este grupo es el que realiza mayores desplazamientos para llegar a la ubicación del establecimiento. Finalmente, los pagados con IDE alto tienen el mejor nivel de educación con índices que van desde 1 a 1,5, es un clúster muy variado en los grupos socio económicos de sus alumnos, pero en promedio se encuentran en el percentil 60. Los estudiantes de esta agrupación ocupan el segundo lugar dentro de los que más se desplazan para llegar a sus establecimientos educacionales.

En el caso de los clústers de matrículas, en ambos casos también se generan 4 clústers. En los resultados de la ejecución solo con atributos propios de las matrículas con una estabilidad inferior o igual al 80 % hay varios factores que generan diferencias, siendo uno de ellos el género. Debido a esto se forman 2 grupos de hombres y 2 de mujeres. Los clústers de

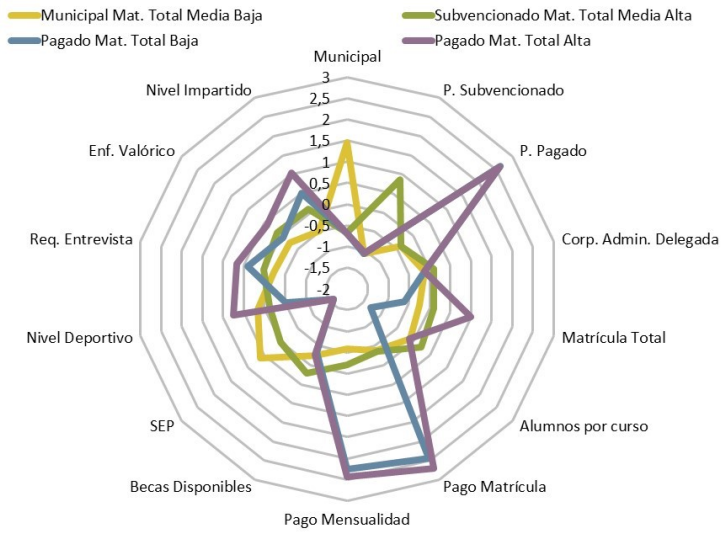


Figura 4.1: Promedios de atributos normalizados de establecimientos de la Región Metropolitana.

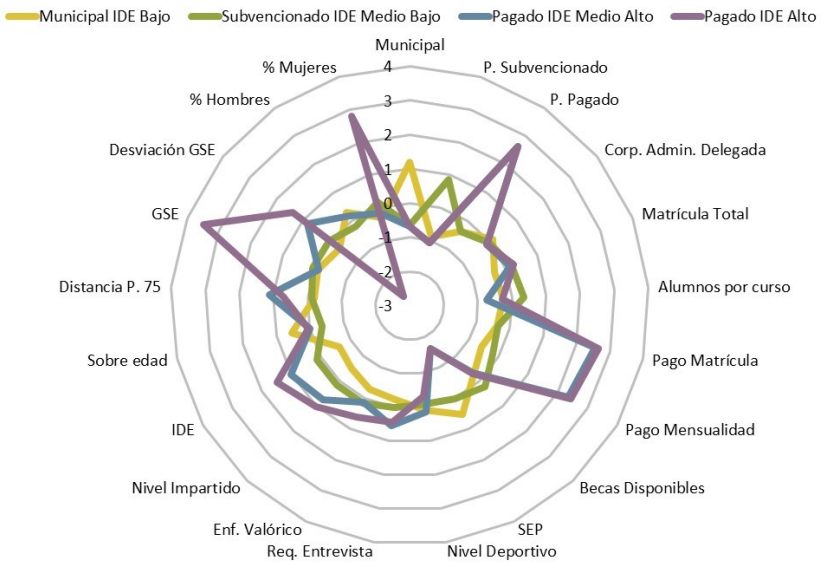


Figura 4.2: Promedios de atributos normalizados de establecimientos con atributos relacionales de la Región Metropolitana.

hombres se diferencian por poseer o no la subvención escolar preferencial, por su nivel de sobre edad y por el grupo socio económico. El primero de ellos (Hombres Sin Beneficio) agrupa a alumnos de un GSE más bajo y con un IPAE inferior, por lo cual se puede decir que este grupo representa a alumnos de familias más pobres y que no viven en los mejores sectores, sectores con bajo acceso a servicios públicos y lugares recreacionales. Es el clúster con el índice de sobre edad mayor, lo que indica que los alumnos que a él pertenecen son los que van más atrasados, es decir, tienen más edad que la que indicada para el nivel que cursan. Por otro lado, el otro grupo de hombres en su mayoría no posee el beneficio SEP, viven en lugares con mejores accesos y son en promedio de un grupo socio económico más alto. Otro punto de comparación entre ambos, y que se puede apreciar en la tabla 4.6, es la cardinalidad de estos, en donde el primero tiene más de 5 veces el número de alumnos que posee el segundo.

Con los clústers de mujeres ocurre una situación similar a la de los hombres, en donde un grupo tiene mejores condiciones que el otro. El grupo de Mujeres Sin Beneficio esta formado por alumnas de familias de un grupo socio económico superior, que viven en mejores sectores y tienen un menor índice de repitencia. A diferencia de este, el grupo de Mujeres Con Beneficio pertenece en promedio a un GSE inferior (percentil 40), viven en peores sectores, con menos accesos a diferentes servicios y tienen un nivel de sobre edad mayor. Además, representan a gran parte de las escolares, debido a que superan en más de 5 veces el número de mujeres del otro grupo.

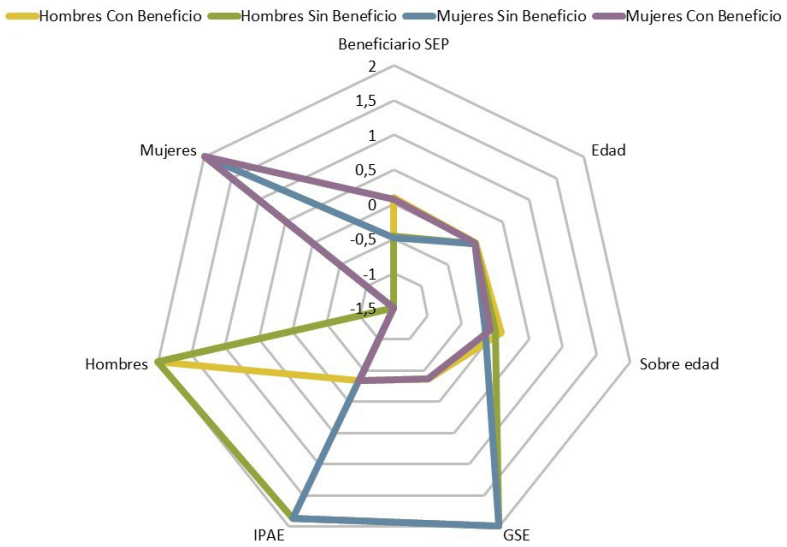


Figura 4.3: Promedio de atributos normalizados de matrículas de la Región Metropolitana.

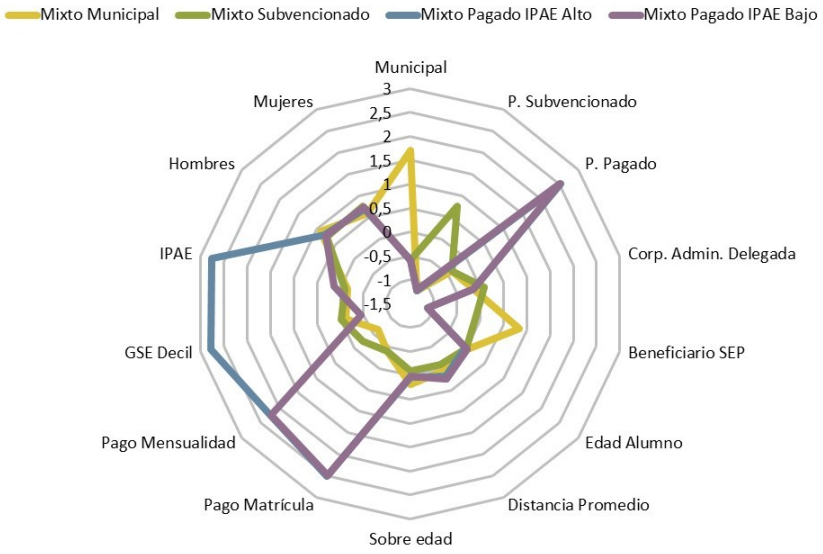


Figura 4.4: Promedio de atributos normalizados de matrículas con atributos relacionales de la Región Metropolitana.

A la hora de incluir los atributos de relación (dependencia del establecimiento, nivel de copago y distancia al establecimiento), los atributos género y el ser o no beneficiario del SEP dejan de ser los principales al momento de describir el clúster y son reemplazados

por la dependencia y el nivel de copago, ya que en esta ejecución los grupos generados no discriminan por sexo. Es decir, en este caso todos los grupos de estudiantes obtenidos son mixtos.

El primer clúster esta formado por estudiantes que asisten a colegios municipales, por lo que sus matrículas y mensualidades son gratuitas, en promedio pertenecen a un GSE bajo (percentil 40), tienen el mayor índice de repitencia y están ubicados en la segunda posición de los que menos se desplazan para asistir a clases. El mixto subvencionado esta compuesto por alumnos de establecimientos particulares subvencionados y de corporación de administración delegada, con un nivel de copago bajo y un nivel socio económico medio (percentil 50). Estos estudiantes presentan el nivel de repitencia más bajo de los 4 grupos, pero se debe considerar que este es el grupo más numeroso, el cual concentra a más de 600.000 alumnos y son los que viven más cerca de sus colegios.

Por último los clúster tercero y cuarto presentan muchas características en común, como estar formados por alumnos de colegios particulares pagados de costos elevado (matrículas sobre $50.000 y mensualidades sobre $100.000) que no poseen el beneficio SEP y se desplazan más de 5 kilómetros para llegar a su lugar de estudio. La diferencia que existe entre ambos radica en el GSE al que pertenecen y al sector donde viven, los estudiantes del tercer clúster viven en lugares que presentan mejores accesos a diferentes servicios y son de un grupo socio económico alto (percentil 90). Por otro lado, en el cuarto clúster los alumnos viven en peores sectores y son de un nivel socio económico bajo (percentil 40).

Realizar una comparación directa entre ambos resultados es una tarea compleja debido a las grandes diferencias que existen en los atributos que permiten describir a cada uno de los clústers, pero podemos notar que si existen algunos factores que con mayor o menor participación en ambos casos generan diferencias entre los grupos, como lo es el caso del GSE. Otro punto a destacar es que en ambos casos el número óptimo de clústers generados por el algoritmo es de 4.

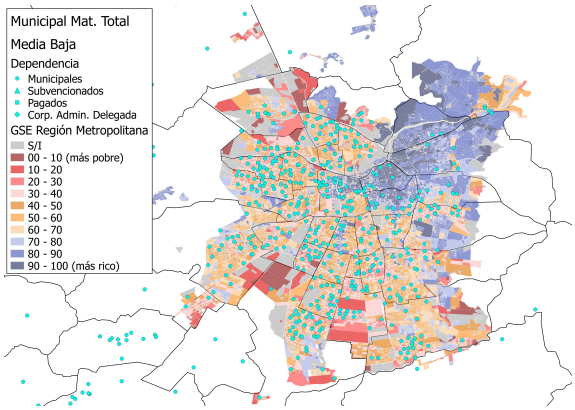
4.2.2. Análisis geográfico

Una vez realizado el análisis cualitativo de los clústers, y con los datos geográficos disponibles, se analiza la distribución geográfica y socioeconómica en el área metropolitana de los clústers de establecimientos, las matrículas de los alumnos correspondientes al clúster del colegio al que asisten y los clústers de las matrículas.

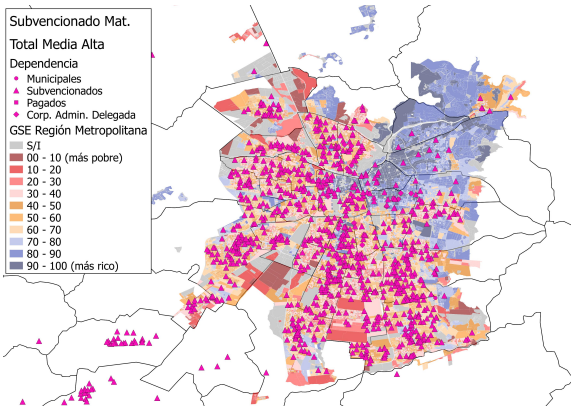
En la figura 4.5 se muestra la distribución de los establecimientos de cada clúster en la Región Metropolitana, incluyendo los grupos socio económicos. Los colegios municipales de matrícula total media baja, como se puede ver en la imagen 4.5a, se encuentran distribuidos de manera uniforme por el área metropolitana, exceptuando la zona nororiente en donde su presencia disminuye notoriamente. Estos se encuentran ubicados en su mayoría en zonas de color anaranjado, las que corresponden a los GSE de los deciles del 3 al 7 y que ocupan gran parte del área metropolitana. Además, esta visión nos permite apreciar de manera sencilla la composición del clúster según la dependencia de sus establecimientos, en donde en este caso predominan los municipales.

Los establecimientos subvencionados de matrícula total media alta en la figura 4.5b se distribuyen de manera similar al clúster anterior, con la diferencia de que por tener una cardinalidad mayor las distancias entre ellos son menores, por lo cual una familia que vive en estos sectores tiene más colegios cerca de este tipo. Estos colegios son en su gran mayoría de dependencia particular subvencionada e incorporando a estos los de corporación de administración delegada como se aprecia en la imagen.

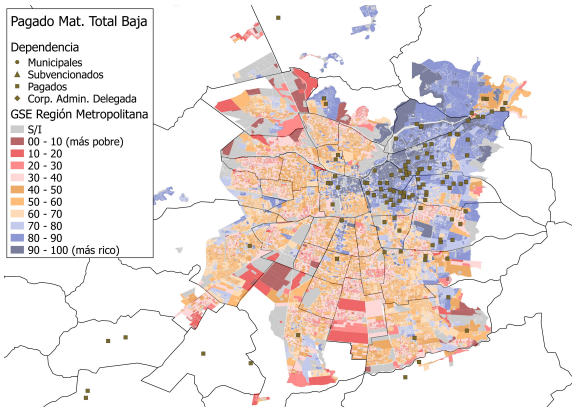
Por otro lado, los establecimientos pagados de matrículas totales bajas y altas, figuras 4.5c y 4.5d respectivamente, se ubican principalmente en zonas de color azul correspondientes a los deciles del 8 al 10 y que mayoritariamente se encuentran en el cono nororiente de la capital, donde viven las familias más acomodadas y con mayor ingreso per cápita. Estos son en su totalidad colegios particulares pagados, que solo se diferencian entre sí por la cantidad de alumnos matriculados, los alumnos por curso y la cantidad de becas que entregan. El pagado de matrículas totales bajas tiene una matrícula total menor, cursos con menos alumnos y menos becas que los pagados de matrículas totales altas.



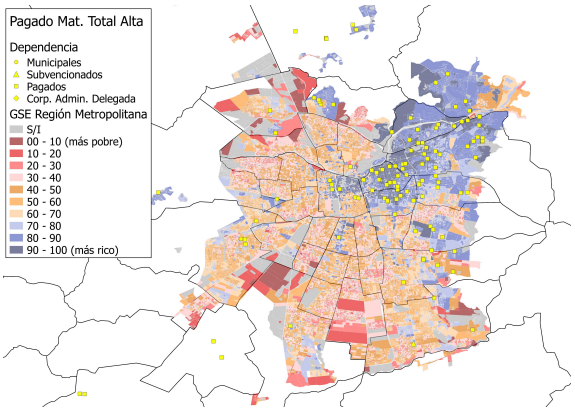
(a) Establecimientos clúster Municipal Mat. Total Media Baja.



(b) Establecimientos clúster Subvencionado Mat. Total Media Alta.



(c) Establecimientos clúster Pagado Mat. Total Baja.

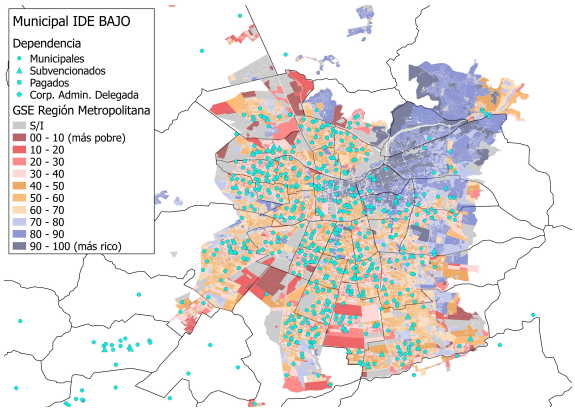


(d) Establecimientos clúster Pagado Mat. Total Alta.

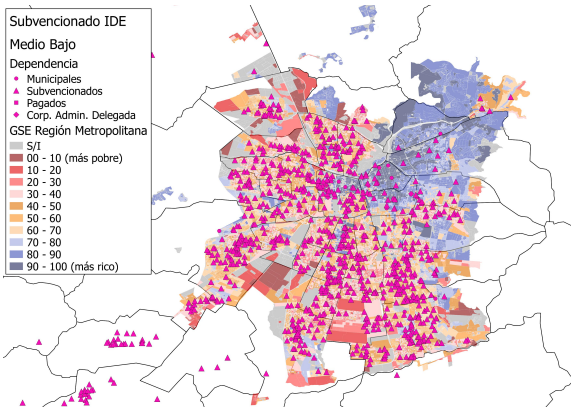
Figura 4.5: Mapas de clústers de establecimientos sobre mapa GSE de la Región Metropolitana.

En el caso de los clústers obtenidos al incluir los atributos de relación establecimiento-matrícula, representados en la figura 4.6, se puede apreciar que la distribución de los primeros dos clústers es similar a los primeros dos clústers descritos para el caso de los clústers generados solo con atributos propios de los establecimientos. Con la diferencia de que esta vez el primer clúster incluye también algunos colegios particulares subvencionados y de corporación de administración delegada. De la misma forma, ahora en el segundo clúster se pueden apreciar algunos colegios municipales, lo que no tiene mayor impacto en la distribución que estos tienen en la capital.

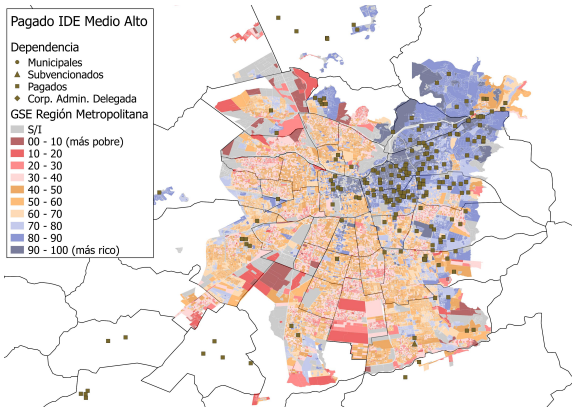
A diferencia de lo ocurrido con los primeros clústers, el tercero y cuarto presentan un mayor cambio, ya que como se puede apreciar en las imágenes 4.6c y 4.6d, el número de colegios aumentó en los pagados de IDE medio alto y disminuyó en los pagados de IDE alto. Esto provoca que en la figura 4.6c los establecimientos se encuentren más agrupados, es decir la distancia entre ellos disminuye y en la figura 4.6d las distancias aumenten.



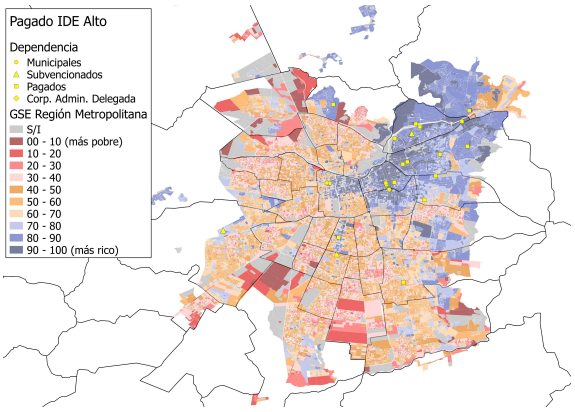
(a) Establecimientos clúster Municipal IDE Bajo.



(b) Establecimientos clúster Subvencionado IDE Medio Bajo.



(c) Establecimientos clúster Pagado IDE Medio Alto.



(d) Establecimientos clúster Pagado IDE Alto.

Figura 4.6: Mapas de clústers de establecimientos (con atributos relacionales) sobre mapa GSE de la Región Metropolitana.

De manera general se puede apreciar que los establecimientos se encuentran principalmente en el área metropolitana de la capital (zonas coloreadas según su GSE). Además, estos se

ubican principalmente en las zonas de colores naranjos y azules, correspondientes a los deciles 3 al 10, dejando a los sectores más pobres (deciles 1 y 2) con una presencia casi nula. Esto, de una u otra manera, refleja la importancia que tiene el grupo socio económico sobre la ubicación de un establecimiento y los niveles de costo que tienen.

En los mapas de la figura 4.7 a la 4.16, utilizando la geolocalización de las matrículas, se muestra la distribución de los alumnos que asisten a los establecimientos de los diferentes clústers. Es decir, en cada mapa se muestran los alumnos pertenecientes a los colegios que conforman los clústers descritos anteriormente. En el mapa 4.7 los alumnos se distribuyen por casi toda el área metropolitana, exceptuando el sector nororiente y concentrándose en sectores de la zona sur y poniente, en comunas como Quinta Normal, Renca, Cerro Navia, San Bernanrdo y El Bosque. Estos se ubican en su gran mayoría en zonas anaranjadas, correspondientes a los deciles del 3 al 7. Al comparar las ubicaciones de los estudiantes con la de los establecimientos se aprecia que se ubican en las mismas zonas.

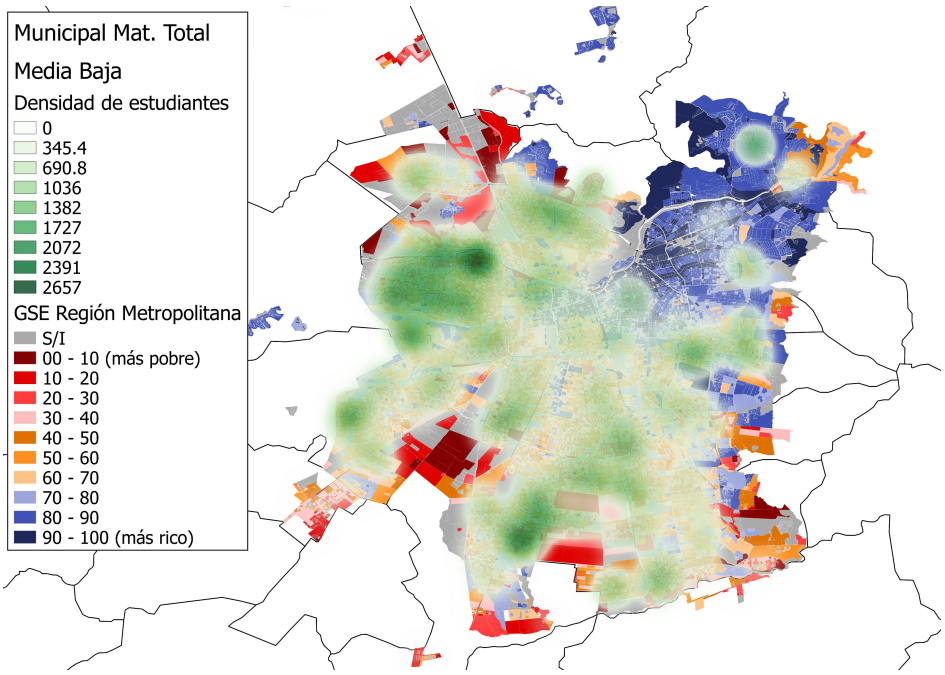


Figura 4.7: Matrículas en colegios del clúster Municipal Mat. Total Media Baja.

El siguiente clúster (figura 4.8), el subvencionado de matrícula total media alta, presenta una distribución similar a la anterior, cubriendo principalmente el sector anaranjado del área

metropolitana y en algunos sectores de la zona oriente. Este presenta sus mayores concentraciones en la zona sur de santiago, como San Bernardo, El Bosque, La Granja, Puente Alto y Maipú. En este caso la ubicación de los alumnos coincide con las ubicaciones de los establecimientos a los que asisten.

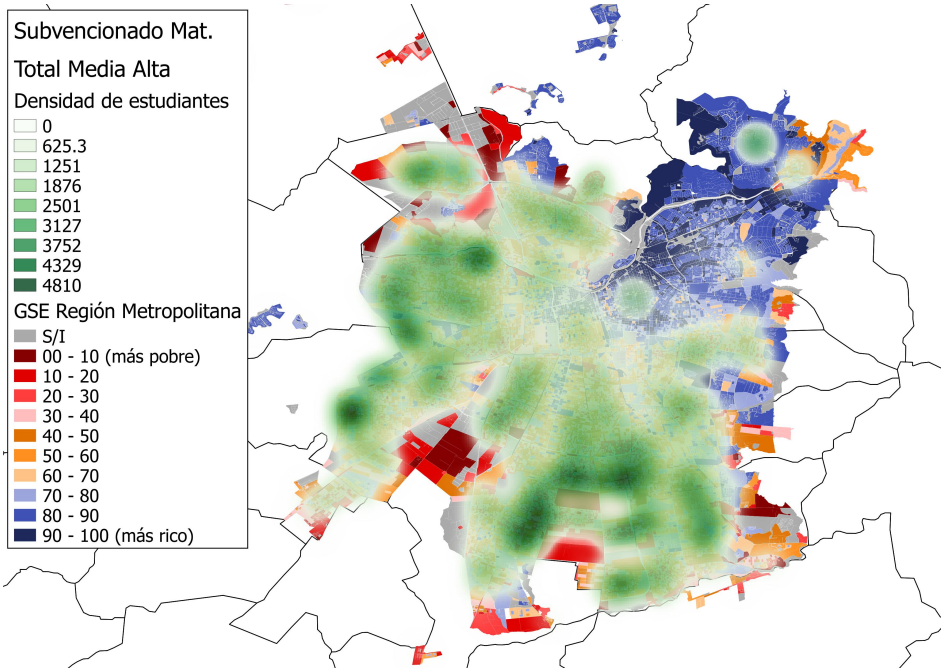


Figura 4.8: Matrículas en colegios del clúster Subvencionado Mat. Total Media Alta.

Los alumnos de los colegios pertenecientes al clúster de los pagados de matrícula total baja (figura 4.9) viven en el sector nororiente de Santiago, pero principalmente en las comunas de Ñuñoa, Providencia, Las Condes, La Reina, Vitacura y Lo Barnechea, lugar donde alcanza su máxima concentración. Los grupos socio económicos en donde residen estos estudiantes varía entre el decil 6 y 10, concentrándose mayoritariamente entre los deciles 8 y 10. Aquí una vez más la ubicación de los establecimientos y el lugar de residencia de los alumnos coincide.

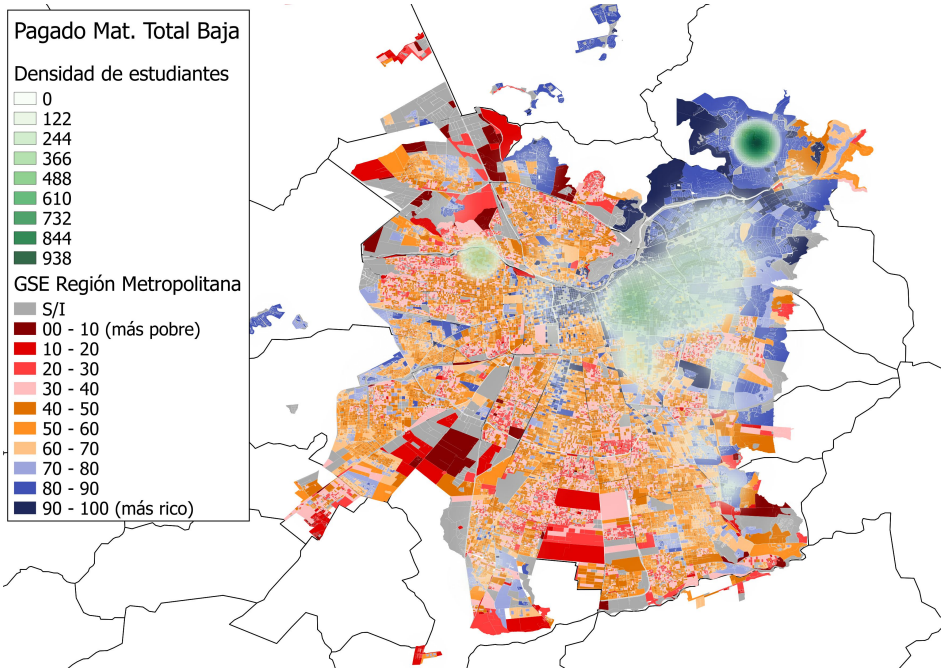


Figura 4.9: Matrículas en colegios del clúster Pagado Mat. Total Baja.

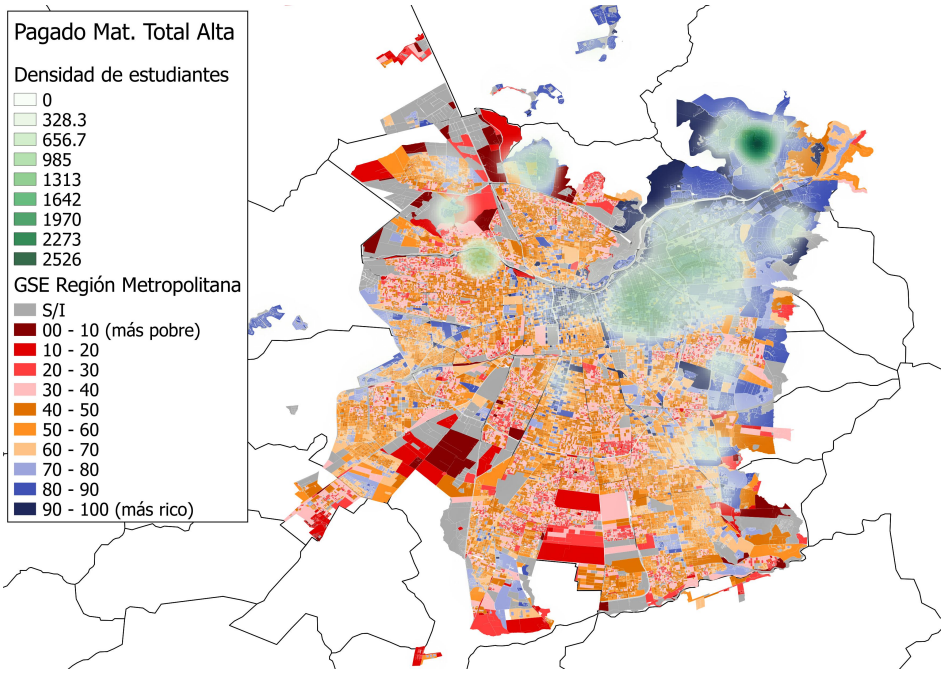
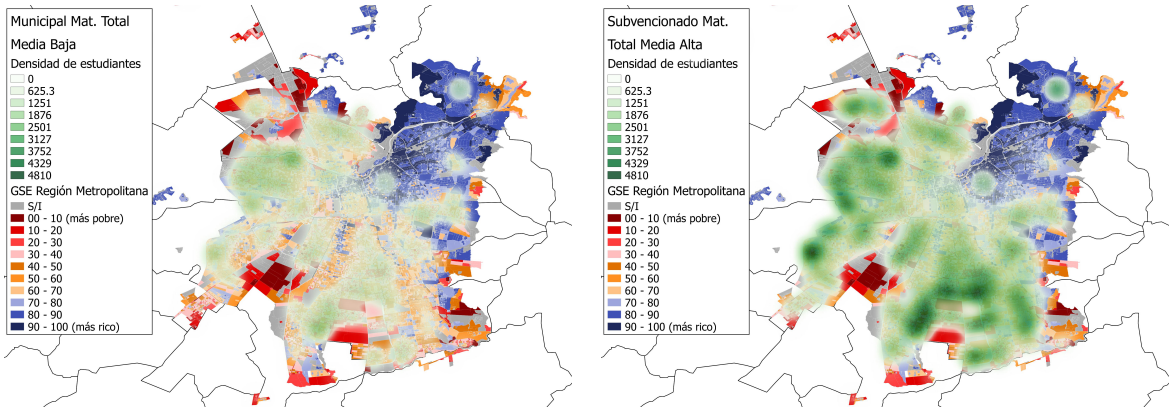


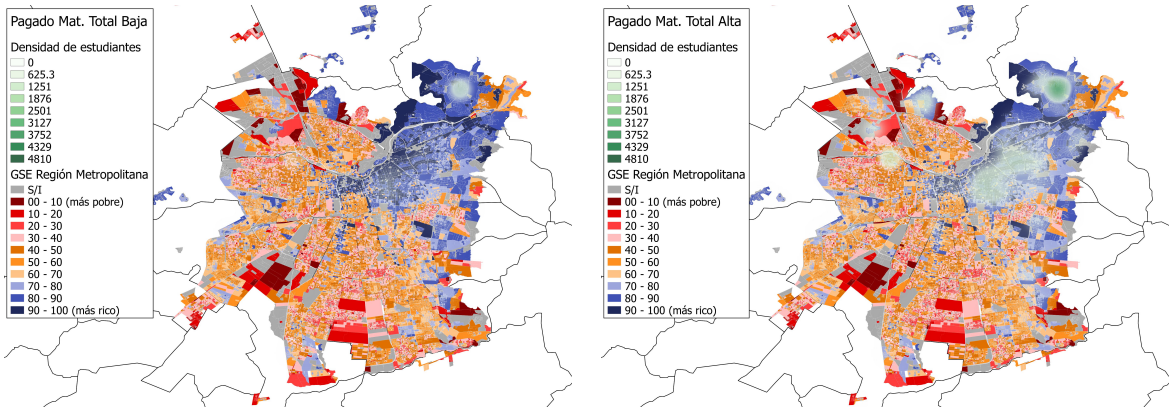
Figura 4.10: Matrículas en colegios del clúster Pagado Mat. Total Alta.

El último clúster de establecimientos generado solo con atributos propios (figura 4.10) tiene sus alumnos ubicados en el sector oriente, pero concentrados principalmente en el cono nororiente de la ciudad y en sectores pertenecientes a los deciles del 8 al 10, alcanzando su máxima densidad en la comuna de Lo Barnechea. Es decir, los estudiantes de los colegios pertenecientes a este clúster viven en los lugares más acomodados de la ciudad.



(a) Municipal Mat. Total Media Baja.

(b) Subvencionado Mat. Total Media Alta.



(c) Pagado Mat. Total Baja.

(d) Pagado Mat. Total Alta.

Figura 4.11: Mapas de calor de matrículas en clústers de establecimientos sobre mapa GSE de la Región Metropolitana.

Ya analizado cada clúster por separado, es momento de realizar una comparación entre ellos, para lo cual en la figura 4.11 se utilizó una escala común que permite apreciar a simple vista las diferencias y semejanzas que entre ellos existen. Lo primero que se aprecia es que la distribución de los primeros dos son muy similares, esparciéndose por gran parte de la capital,

exceptuando el cono nororiente. La diferencia que existe entre ambos está dada por la cantidad de estudiantes que pertenecen a cada uno de ellos, en donde el primero agrupa a 124.951 estudiantes y el segundo a 307.459.[1] De manera similar los mapas inferiores (figuras 4.11c y 4.11d) también presentan una distribución semejante, ocupando principalmente el sector nororiente del área metropolitana, el cual corresponde a los grupos socio económicos más altos. En ambos casos la nube de estudiantes es menos notaria en comparación al resto, debido a que el número de estudiantes en ellos es mucho menor (8.501 y 30.555 respectivamente).

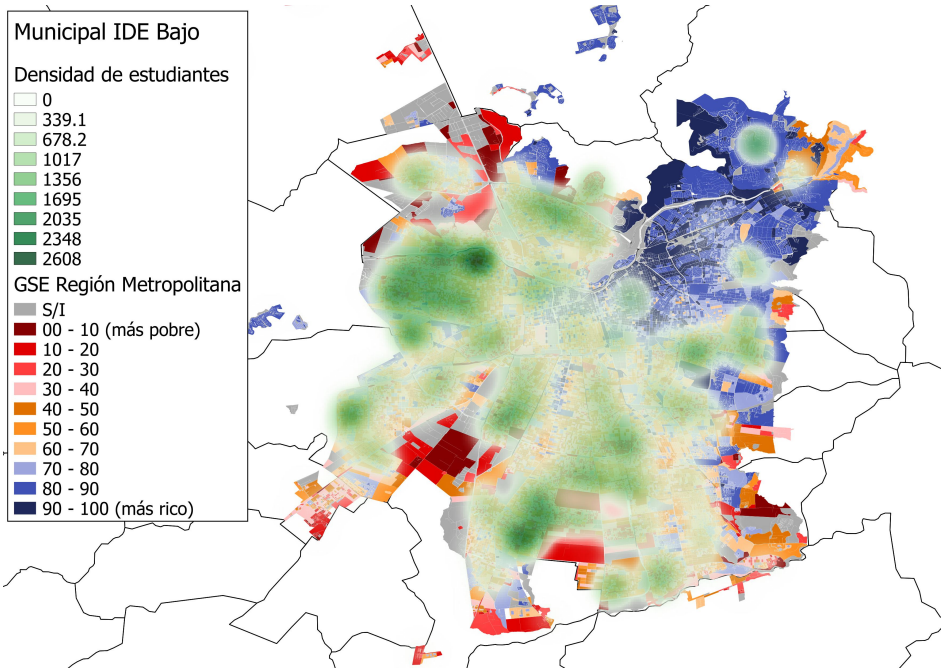


Figura 4.12: Matrículas en colegios del clúster Municipal IDE Bajo.

Los clústers obtenidos al incorporar los atributos de relación, y como se mencionó anteriormente, presentan pequeñas diferencias con respecto al primer análisis sin dichos atributos. En el caso de los primeros dos clústers (figuras 4.12 y 4.13) no es posible observar dichos cambios a nivel visual, debido a que el número de estudiantes es similar y no afectó de manera significativa la densidad de estudiantes en los diferentes sectores. Las matrículas del clúster de establecimientos municipales de IDE bajo se sitúan en las zonas anaranjadas (deciles 3 al 7), concentrándose en las comunas de Quinta Normal, Cerro Navia, San Bernardo y El Bosque. De igual forma lo hacen los estudiantes de los establecimientos subvencionados

[1] Las cifras corresponden a los estudiantes pertenecientes al clúster que poseen geolocalización.

de IDE medio bajo, que se concentran en las comunas de Maipú, San Bernardo, El Bosque, La Granja y Puente Alto.

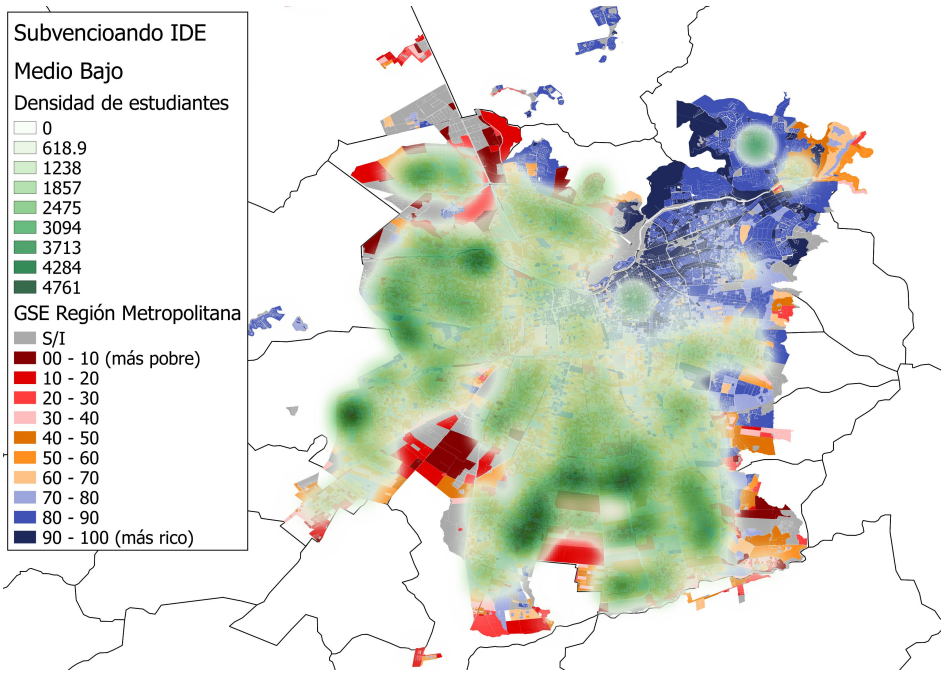


Figura 4.13: Matrículas en colegios del clúster Subvencionado IDE Medio Bajo.

Para los colegios pertenecientes al clúster de pagados de IDE medio alto (figura 4.14), sus estudiantes viven mayoritariamente en los sectores correspondientes a grupos socio económicos entre el decil 8 y 10, mismo lugar donde se concentran los establecimientos de dicho clúster. Estos agrupan a 34.298 estudiantes, los cuales alcanzan su mayor concentración (densidad de 3.112 estudiantes) en la comuna de Lo Barnechea.

En la figura 4.15 los estudiantes pertenecientes a dicho clúster se concentran principalmente en la comuna de Maipú y en menor medida en las comunas de Providencia, Las Condes y Lo Barnechea. Mismos lugares en donde se ubican los 23 establecimientos que conforman el grupo, número que explicaría la baja densidad que este presenta.

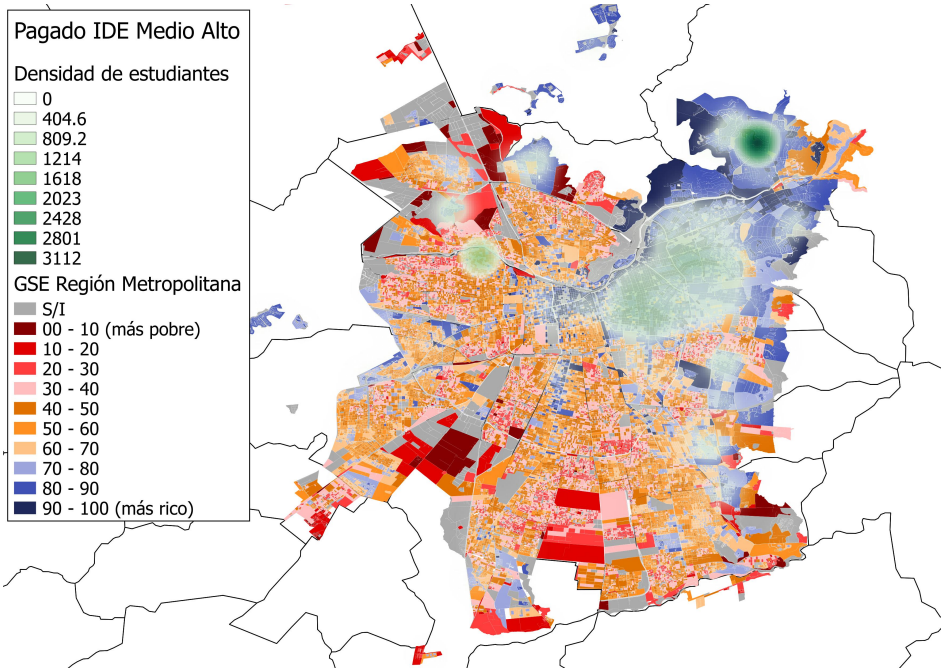


Figura 4.14: Matrículas en colegios del clúster Pagado IDE Medio Alto.

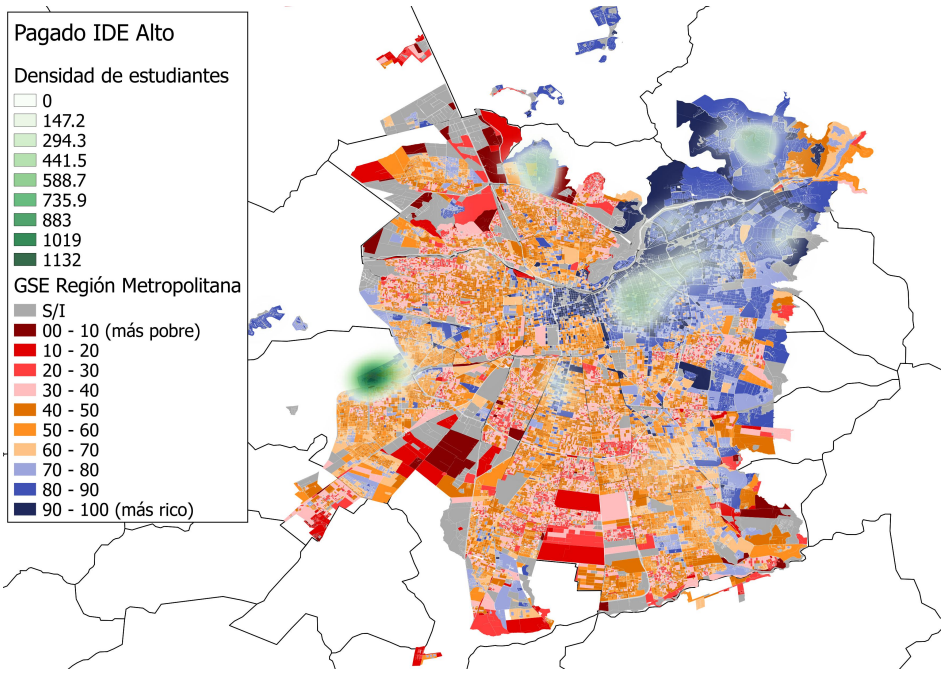
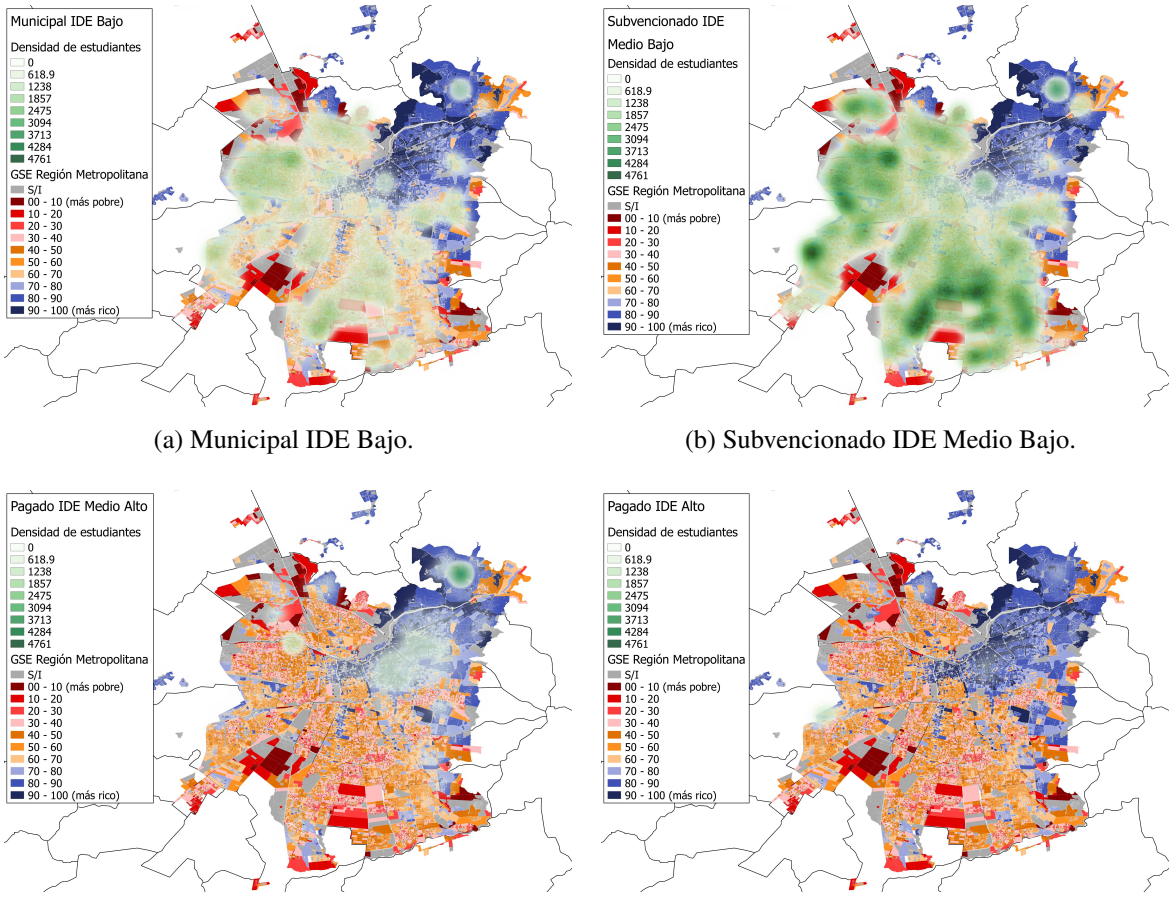


Figura 4.15: Matrículas en colegios del clúster Pagado IDE Alto.

Al utilizar la misma escala de densidad de estudiantes, como se muestra en la figura 4.16,

se puede ver la diferencia en las ubicaciones de los estudiantes, las cuales parecen ser complementarias. En dos de las figuras los estudiantes viven en las zonas anaranjadas, que son aproximadamente tres cuartos del área metropolitana y los otros dos cubren los lugares de colores azules. Además se puede apreciar a simple vista cuales son los que presentan una mayor cantidad de alumnos, como lo es el que se representa en la figura 4.16b con 307.066 matriculados y los que tienen menos, como es el caso de los pagados con IDE alto de la figura 4.16d donde la densidad es muy baja y es difícil de representar con la escala utilizada. El número exacto de estudiantes que pertenece a cada clúster de colegios es de 307.066, 123.126, 34.298 y 7.066, respectivamente.



(a) Municipal IDE Bajo.

(b) Subvencionado IDE Medio Bajo.

(c) Pagado IDE Medio Alto.

(d) Pagado IDE Alto.

Figura 4.16: Mapas de calor de matrículas en clústers de establecimientos (con atributos relacionales) sobre mapa GSE de la Región Metropolitana.

De manera general, en ambos resultados, es posible inferir que en su mayoría los alumnos asisten a establecimientos educacionales que se ubican relativamente cerca o en el mismo sector en el que residen. Es decir, si los colegios del clúster pertenecen a la zona nororiente de Santiago, los alumnos que a ellos asisten viven en el mismo sector o en un sector cercano.

De forma similar a como se analizaron las matrículas en los clústers de establecimientos, se analizan los clústers de matrículas que se obtuvieron solo utilizando las atributos propios de estos y añadiendo los atributos que relacionan la matrícula con el establecimiento al que asiste.

Los estudiantes del clúster hombres con beneficio (figura 4.17) viven en gran parte de la capital, ocupando aproximadamente tres cuartos del área metropolitana. Estos se ubican en los lugares donde predominan los grupos socio económicos medios y bajos (deciles 3 al 7) y se concentran en las comunas de Quinta Normal, San Bernardo y El Bosque. Por otro lado los hombres sin beneficio se ubican a lo largo y ancho de la capital, pero en sectores bien definidos, como los son las comunas de Quilicura, Maipú, Santiago, Puente Alto, La Florida, Ñuñoa, La Reina, Las Condes, Providencia y Lo Barnechea, concentrándose principalmente en las últimas dos. Estos principalmente están situados en sectores donde el decil oscila entre los deciles 6 y 10, es decir, pertenecen a la clase más acomodada.

Los clústers de mujeres sin y con beneficio (figuras 4.19 y 4.20 respectivamente) se comportan de la misma forma en que lo hacen los clústers de hombres sin y con beneficios, en donde los primeros residen en determinados sectores del área metropolitana, predominantemente en lugares donde el GSE es mayor, y los segundos se distribuyen de manera homogénea por la capital, exceptuando el sector nororiente.

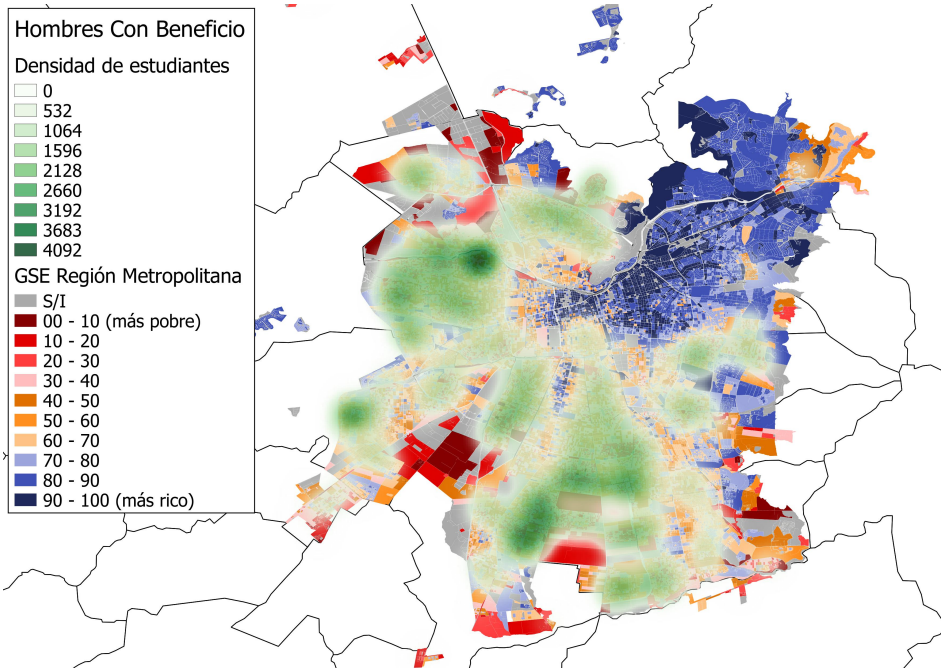


Figura 4.17: Matrículas del clúster Hombres Con Beneficio.

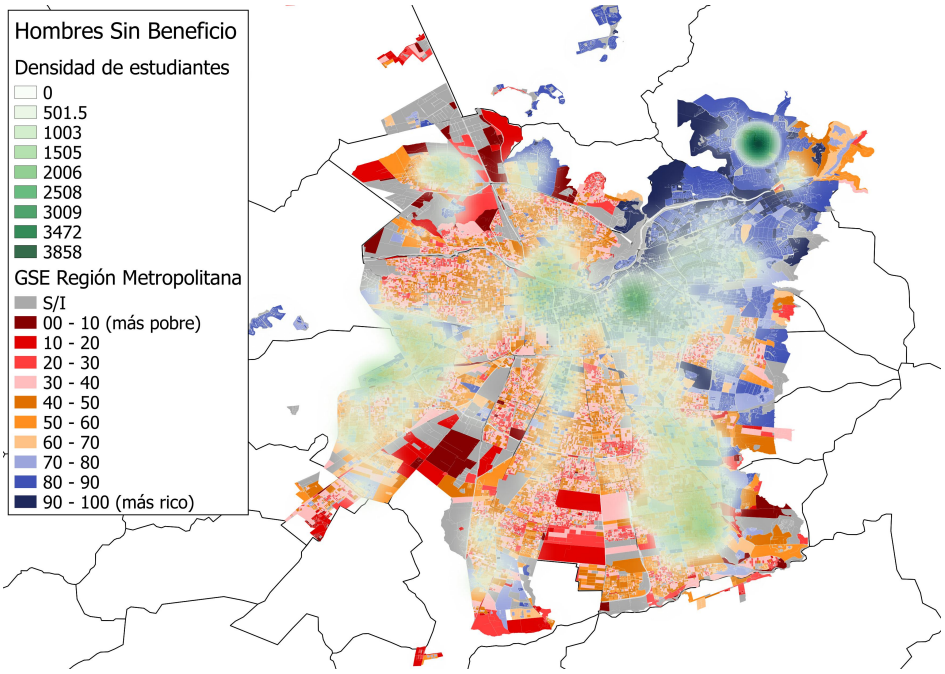


Figura 4.18: Matrículas del clúster Hombres Sin Beneficio.

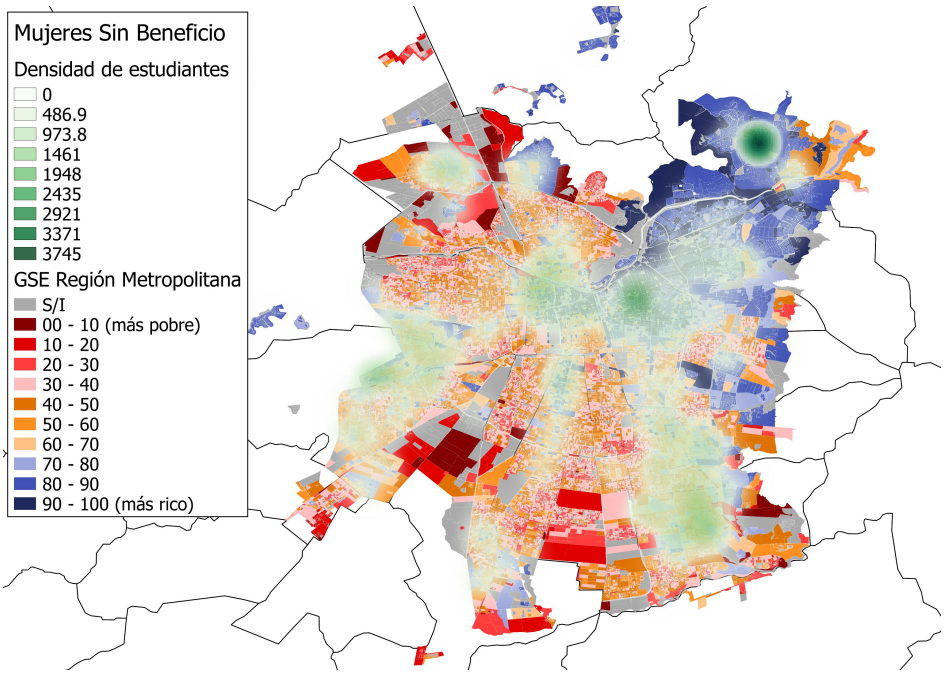


Figura 4.19: Matrículas del clúster Mujeres Sin Beneficio.

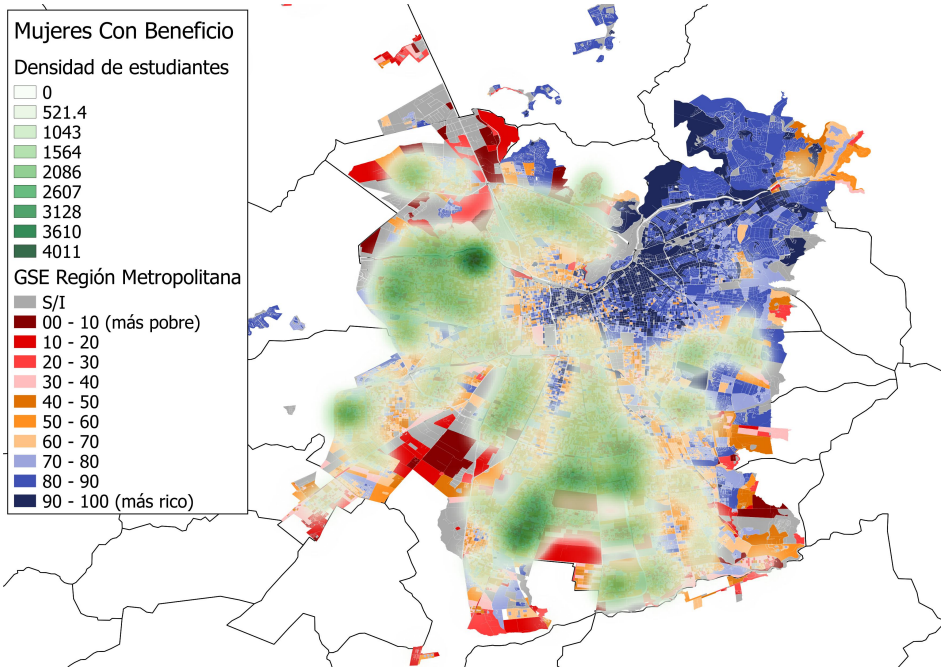
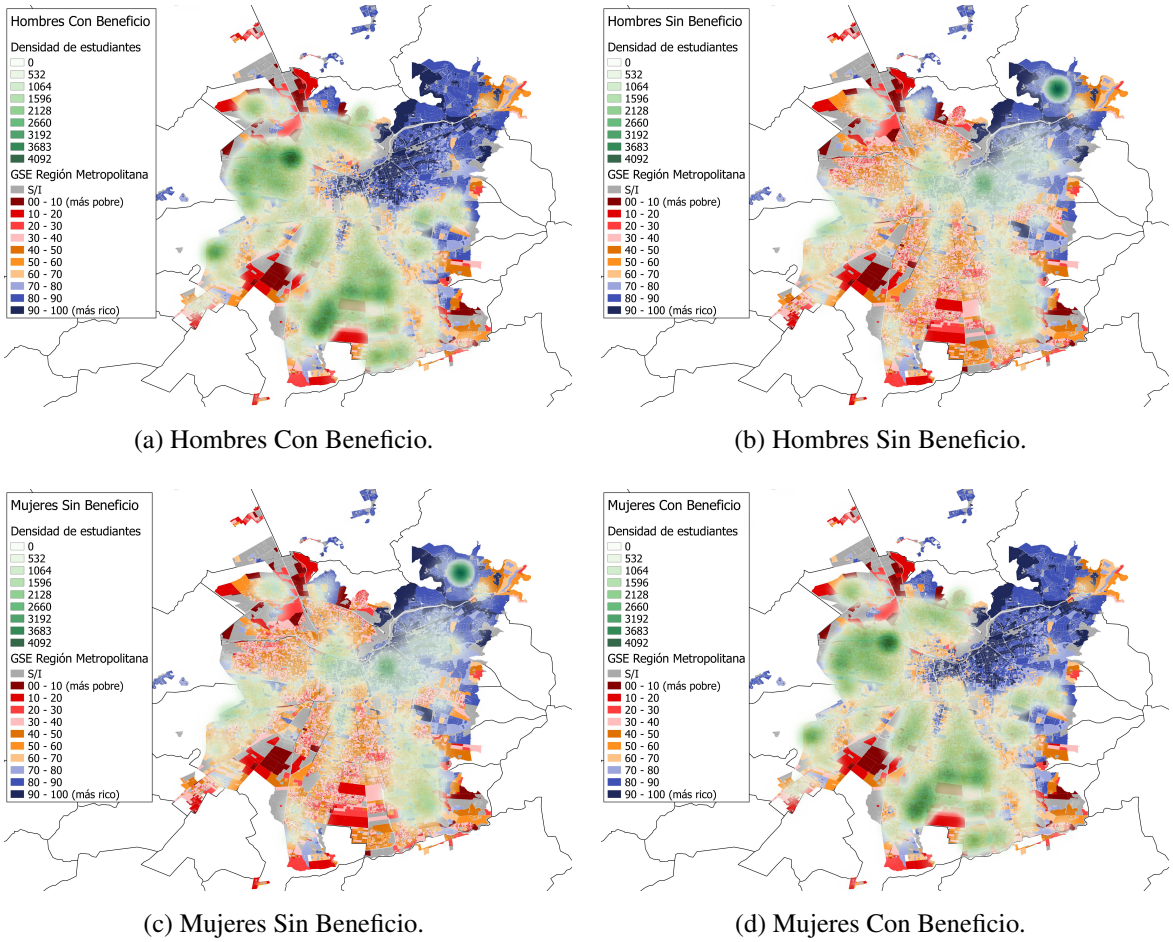


Figura 4.20: Matrículas del clúster Mujeres Con Beneficio.

Al comparar los 4 mapas de los clústers generados para las matrículas con la misma escala de

densidad de estudiantes (figura 4.21), se aprecia de manera clara lo señalado anteriormente, en donde hombres y mujeres con beneficio se distribuyen de manera bastante homogénea por tres cuartas partes de la capital y presentando mayores concentraciones en las comunas de Quinta Normal, El Bosque y San Bernardo. Cabe destacar que estos estudiantes se ubican principalmente en zonas de un ingreso medio, es decir de GSE que van entre el decil 3 al 7.



(a) Hombres Con Beneficio.

(b) Hombres Sin Beneficio.

(c) Mujeres Sin Beneficio.

(d) Mujeres Con Beneficio.

Figura 4.21: Mapas de calor de clústers de matrículas sobre mapa GSE de la Región Metropolitana (misma escala de densidad).

Por otro lado, los hombres y mujeres sin beneficio están ubicados en diferentes comunas de la capital como Maipú, Puente Alto, La Florida, Santiago, Macul, La Reina, Las Condes, Providencia y Lo Barnechea, en donde las últimas dos presentan las mayores concentraciones. Además, esta figura permite apreciar que los clústers con beneficio agrupan una mayor

cantidad de estudiantes, superando cada uno los 150.000, a diferencia de los que no poseen el beneficio, en donde cada uno bordea los 80.000.

Finalmente, se analizan los clústers de matrículas obtenidos al incluir los atributos de relación establecimiento-matrícula. El primero de ellos (figura 4.22) está presente en toda el área metropolitana, pero disminuyendo su densidad en el sector nororiente. Según los lugares en los cuales se encuentra abarca diferentes grupos socio económicos, los cuales van desde uno de los más pobres (decil 3) hasta los más ricos (decil 10). Estos estudiantes asisten en su mayoría a establecimientos municipales, que como se analizó anteriormente, están presentes en gran parte de Santiago.

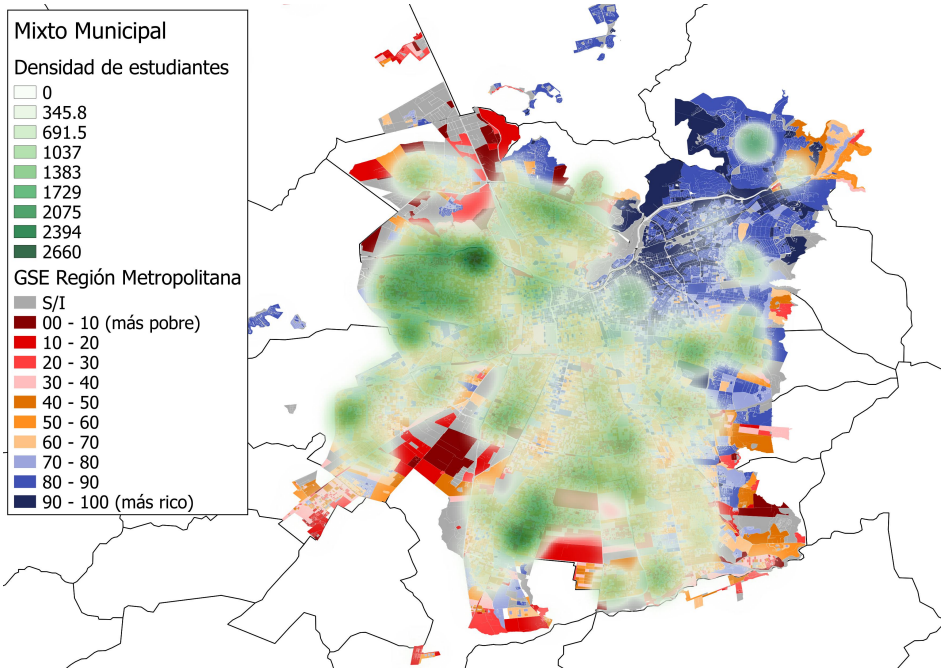


Figura 4.22: Matrículas del clúster Mixto Municipal.

El clúster de matrículas mixto subvencionado (figura 4.23), al igual que el clúster anterior, tiene sus estudiantes repartidos por gran parte de la capital y especialmente en sectores socio económicos medios. Debido a que este agrupa más del doble de matrículas que el clúster anterior y sobre el mismo terreno, ocurre que existen mayores zonas donde se concentran. En este caso las mayores concentraciones se dan en la zona sur (San Bernardo, El Bosque, La Pintana, La Granja y Puente Alto) y en la zona poniente (Maipú, Cerro Navia, Pudahuel,

Lo Prado y Quinta Normal). Al hacer la comparación con la ubicación de los colegios a los que asisten, coincide en que estos se sitúan en la misma zona donde viven los estudiantes.

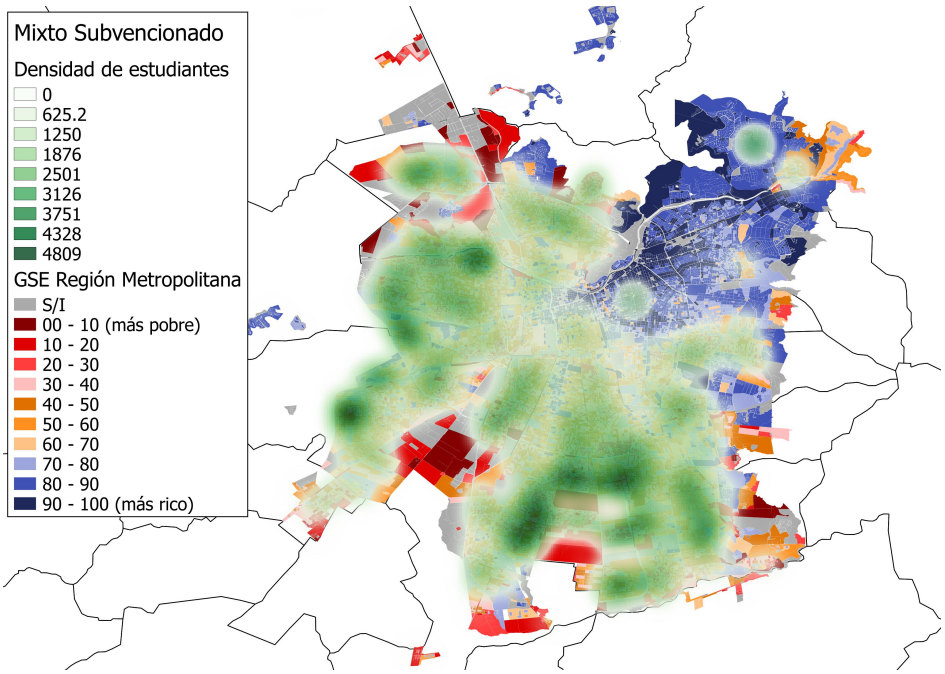


Figura 4.23: Matrículas del clúster Mixto Subvencionado.

Las matrículas del clúster mixto pagado de IPAE alto (figura 4.24) residen en zonas correspondientes a los grupos socio económicos más adinerados, los que en su mayoría se ubican en la zona nororiente de la capital, lo que además concuerda con que su nivel IPAE sea de los más altos. Al igual que el resto de los casos, existe un lugar donde la densidad de estudiantes es mayor y en este clúster esto ocurre en la comuna de Lo Barnechea. Otro punto importante es que en este caso también coincide la ubicación de los colegios con la de los alumnos, es decir se encuentran en los mismos sectores.

Por último el cuarto clúster, representado en la figura 4.25, agrupa la mayor cantidad de sus alumnos en la comuna de Quinta Normal y es el único caso en que la ubicación de los establecimientos no coincide con la de los alumnos, lo cual lleva a entender el por qué son los que más se desplazan para llegar a sus establecimientos educacionales.

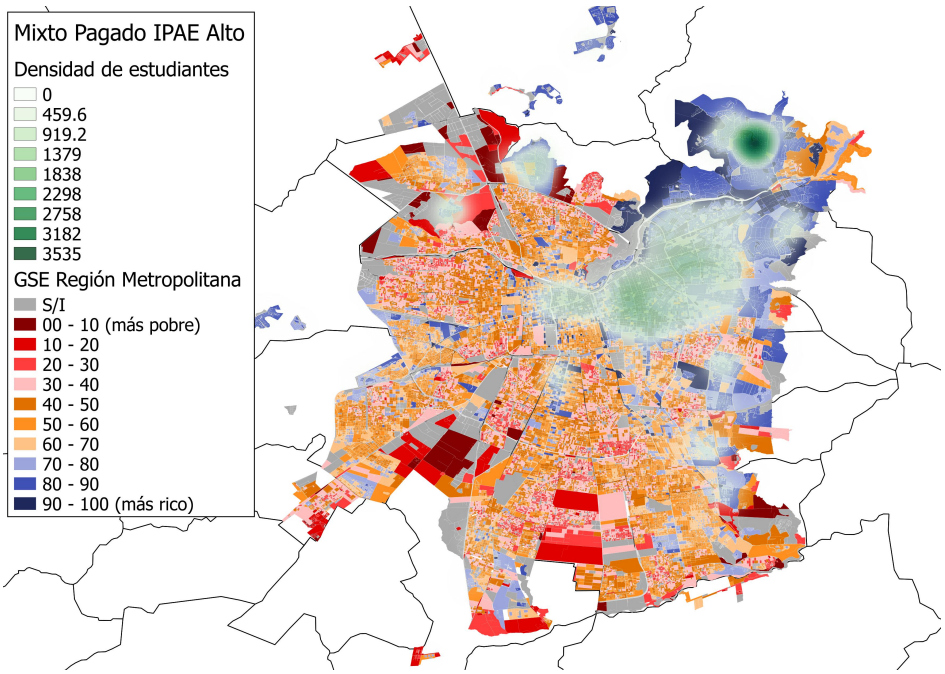


Figura 4.24: Matrículas del clúster Mixto Pagado IPAE Alto.

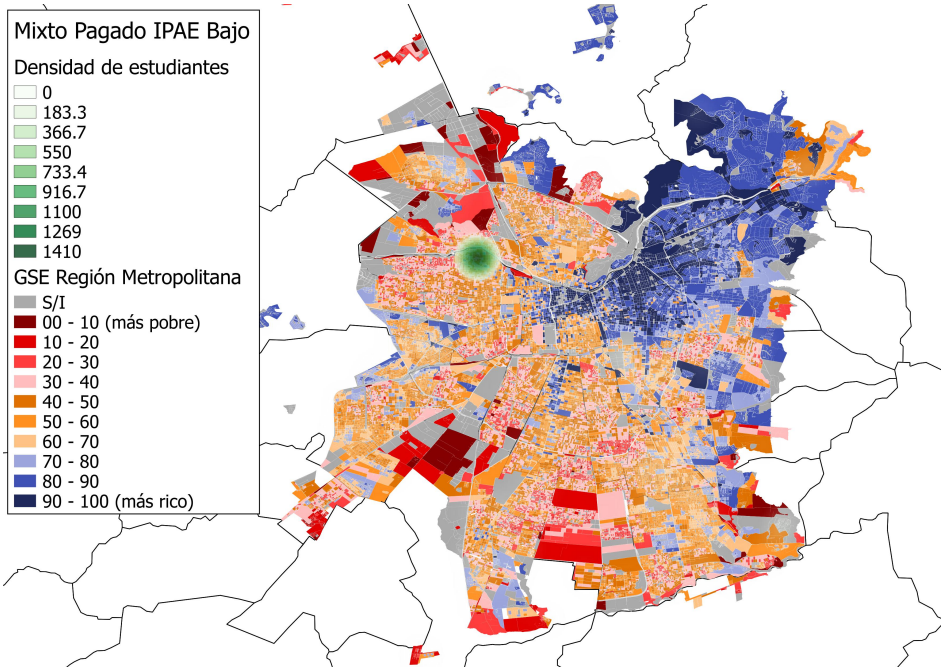
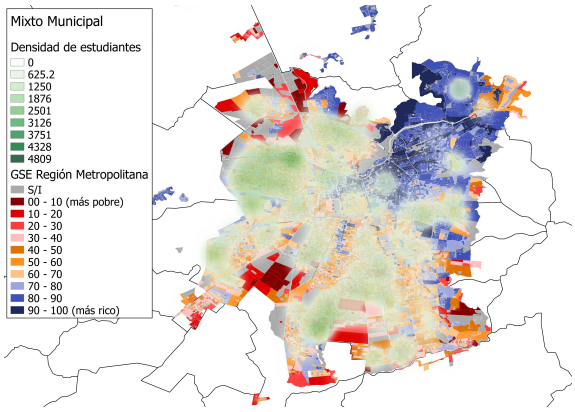


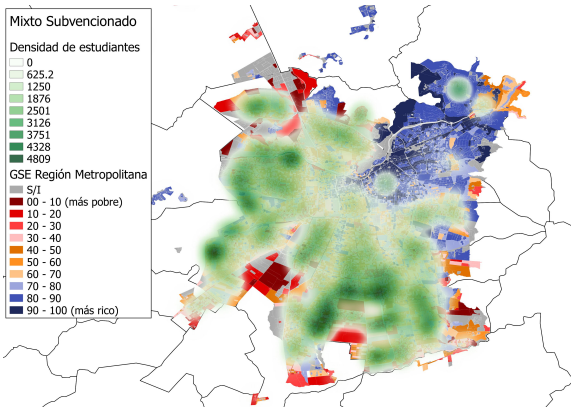
Figura 4.25: Matrículas del clúster Mixto Pagado IPAE Bajo.

En la figura 4.26 se muestran los mapas con los clústers de matrículas con una escala de

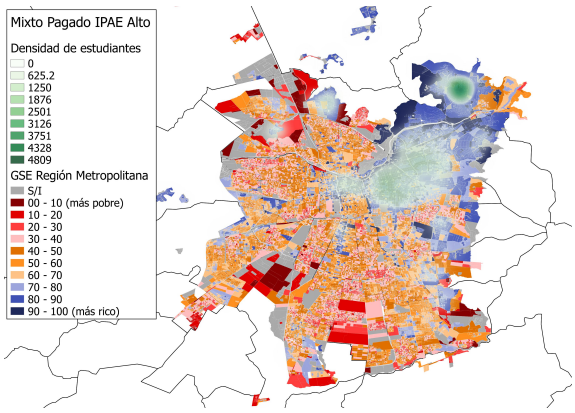
densidad semejante, para poder realizar de mejor manera una comparación entre ellos. En estas imágenes se puede inferir con claridad cuales son los clústers que más estudiantes agrupan y los que menos. Si se superponen todos los mapas de calor, se apreciaría la densidad general de los estudiantes en el área metropolitana. En donde se mostrarían los sectores de la capital donde viven más estudiantes y que por lo general coincide con los lugares más poblados [12], como lo son las comunas de Santiago, Puente Alto, San Bernardo, El Bosque y La Granja entre otras.



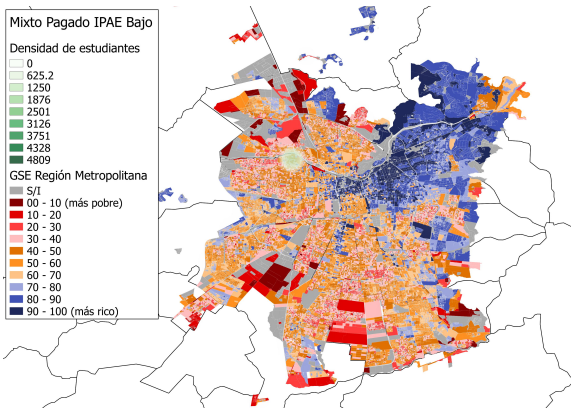
(a) Mixto Municipal.



(b) Mixto Subvencionado.



(c) Mixto Pagado IPAE Alto.



(d) Mixto Pagado IPAE Bajo.

Figura 4.26: Mapas de calor de clústers de matrículas (con atributos relacionales) sobre mapa GSE de la Región Metropolitana (misma escala de densidad).

Conclusiones

En este último capítulo se presentan las conclusiones finales de la investigación, posibles extensiones y el cumplimiento de los objetivos propuestos en la definición del problema.

El estudio consistió principalmente en la realización de experimentos sobre bases de datos de establecimientos y matrículas de la Región Metropolitana, con atributos propios de cada uno y atributos que relacionan a ambas entidades, para encontrar los grupos que mejor modelen la realidad de nuestra capital.

Para la ejecución del estudio se utilizó en primera instancia R para el manejo de los datos y luego RapidMiner para llevar a cabo los experimentos. Como uno de los objetivos era que no hubiera un sesgo inicial de como debían ser los grupos, el algoritmo escogido fue el de X-Means, el cual corresponde a un algoritmo de clusterización no supervisado.

En el desarrollo de este experimento se utilizó la metodología CRISP-DM, la que es una de las más utilizadas para proyectos de Data Science, la cual permitió llevar a cabo de buena manera el estudio y cumplir los objetivos planteados para este.

La experiencia consistió principalmente en anotar y describir los resultados que se obtienen al ejecutar X-Means sobre los atributos propios de los establecimientos y de las matrículas, para luego realizar una nueva ejecución incorporando los atributos que relacionan ambas entidades y poder comparar ambos resultados.

En el caso de los establecimientos los resultados obtenidos en ambas ejecuciones son bastante similares, con la diferencia de que al incluir los nuevos atributos permiten describir de manera mas detallada el clúster, incorporando información importante sobre los alumnos

que estudian en dichos colegios. Además, al realizar los clústers se puede apreciar que hay características que tienen una mayor importancia y otras que no son tan relevantes. Los atributos propios más relevantes son la dependencia del establecimiento, el nivel de copago y la cantidad de alumnos matriculados. De la misma forma, los atributos de relación que más se destacan son el IDE y la distancia a la que viven los estudiantes.

A grandes rasgos, los clústers de establecimientos se encuentran divididos en 4 tipos, los que vendrían siendo: los municipales, los particulares subvencionados y dos grupos de particulares pagados. Estos además de agruparse por la dependencia, se dividen por los niveles de copago que presentan, los que van de colegios gratuitos hasta los que cuestan por sobre los $100.000 pesos chilenos.

Para las matrículas se presenta un escenario más complejo, debido a que los grupos que se obtuvieron en ambas ejecuciones difieren mucho entre ellos. En la primera ejecución las matrículas se dividen por el género, su grupo socio económico, el IPAE y por si tienen o no el beneficio SEP, es decir, por casi todos los atributos propios de dicha entidad. Lo anteriormente descrito cambia drásticamente al incorporar los atributos de relación, ya que los nuevos resultados se ven fuertemente influenciados por el tipo de colegio al que asisten y el nivel de copago que tienen. En este sentido es difícil establecer si un resultado es mejor o peor que el otro y será necesario acompañarlo del enfoque para el cual se quiera utilizar dichos clúster para poder escoger el más indicado.

De los análisis geográficos realizados se puede establecer que los establecimientos se ubican en sectores acordes a su nivel de copago, es decir, los colegios con niveles de matrículas y mensualidades más bajas se encuentran en las zonas más pobres o de un menor grupo socio económico y los que tienen niveles de copago más elevado se sitúan en los sectores más acomodados y con mayores ingresos. De manera similar ocurre con los estudiantes, que por lo general viven en barrios de similares características a las que tiene el sector donde estudian. Otra característica importante y que refuerza lo investigado en el estado del arte, es que los estudiantes que asisten a establecimientos particulares pagados están dispuestos a desplazarse más en comparación a alumnos que asisten a colegios municipales. Esto también tiene que ver con la mayor cantidad de colegios municipales y subvencionados que existen en comparación a los particulares pagados, y la distribución de estos en el área metropolitana.

Una recomendación para trabajos a futuro sobre este problema, es considerar obtener o generar una base de datos más robusta, con mayor número de atributos y que cuenten con un nivel de estandarización previa mayor, esto debido a que para este estudio se debieron estandarizar algunos atributos obtenidos de la Web del Ministerio de Educación, quedando a criterio del investigador y del CIAE la estandarización de ciertos atributos.

A modo de conclusión general, la solución planteada y los resultados obtenidos con esta satisfacen los objetivos planteados para el problema, entregando los diferentes clústers y atributos más relevantes para los establecimientos y matrículas de la Región Metropolitana. Estos resultados permitirán a futuro poder definir de mejor manera el público objetivo para el desarrollo de políticas públicas.

Bibliografía

[1] Atila Abdulkadiroğlu and Tayfun Sönmez. School choice: A mechanism design approach. *American economic review*, 93(3):729–747, 2003.

[2] Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[3] Manuel Canales, Cristián Bellei, and Víctor Orellana. ¿Por qué elegir una escuela particular subvencionada? sectores medios emergentes y elección de escuela en un sistema de mercado. *Estudios Pedagógicos (Valdivia)*, 42:89 – 109, 2016.

[4] Yan Chen and Tayfun Sönmez. School choice: an experimental study. *Journal of Economic theory*, 127(1):202–231, 2006.

[5] Rómulo A. Chumacero, Daniel Gómez, and Ricardo D. Paredes. I would walk 500 miles (if it paid): Vouchers and school choice in Chile. *Economics of Education Review*, 30(5):1103 – 1114, 2011.

[6] Ministerio de Chile. MIME - Ministerio de Educación de Chile. `http://www.mime.mineduc.cl/mvc/mime/portada`. Accedido: 7 Oct. 2017.

[7] Ministerio de Desarrollo Social. Observatorio Social - Ministerio de Desarrollo Social - Gobierno de Chile. `http://observatorio.ministeriodesarrollosocial.gob.cl/casen/basededatos_historico.php`. Accedido: 22 Feb. 2018.

[8] David Gale and Lloyd S Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

[9] Francisco A. Gallego and Andrés Hernando. School choice in Chile: Looking at the demand side. `http://economia.uc.cl/publicacion/school-choice-in-chile-looking-at-the-demand-side/`, 09 2009.

[10] Daniel Gómez, Rómulo A. Chumacero, and Ricardo D. Paredes. School choice and information. *Estudios de economía*, 39:143 – 157, 12 2012.

[11] Judith Hurwitz and Daniel Kirsch. *Machine Learning for dummies*. John Wiley & Sons, Inc, 2018.

[12] INE. El nuevo mapa de la población en Chile: Revisa los datos del Censo comuna por comuna. `http://www.emol.com/noticias/Economia/2017/12/22/888510/El-nuevo-mapa-de-la-poblacion-en-Chile-Revisa-los-datos-del-/Censo-comuna-por-comuna.html`. Accedido: 26 Abr. 2018.

[13] Leonard Kaufman and Peter Rousseeuw. *Finding Groups in Data: An Introduction To Cluster Analysis*. John Wiley & Sons, Inc, 2009.

[14] Stuart Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, March 1982.

[15] Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz, and Timm Euler. Yale: Rapid prototyping for complex data mining tasks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 935–940. ACM, 2006.

[16] Ryan Mitchell. *Web scraping with Python: collecting data from the modern web*. O'Reilly Media, Inc., 2015.

[17] Alberto Montresor and Alessio Guerrieri. *Decentralized clustering with estimation of the number of clusters*. Tesis para optar al grado de Máster en Ciencias en Informática, Universidad de Trento, 2010.

[18] Dau Pelleg and Andrew Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *In Proceedings of the 17th International Conf. on Machine Learning*, pages 727–734. Morgan Kaufmann, 2000.

[19] Patrick Royston, Ian R White, et al. Multiple imputation by chained equations (mice): implementation in stata. *J Stat Softw*, 45(4):1–20, 2011.

[20] Claudio Sapelli and Arístides Torche. Subsidios al alumno o a la escuela: efectos sobre la elección de colegios públicos. *Cuadernos de economía*, 39:175 – 202, 08 2002.

[21] Carlos Soto, Ximena Saavedra, Ignacio Larraguibel, and Felipe Flores. Estadísticas de la educación 2016. page 15, 2017.

[22] Rüdiger Wirth. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, pages 29–39, 2000.