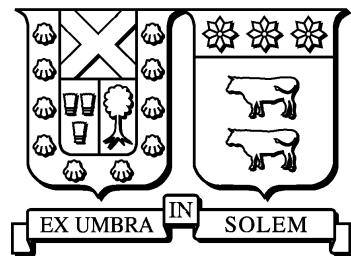


UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE INFORMÁTICA
SANTIAGO – CHILE



“TÍTULO DE LA MEMORIA”

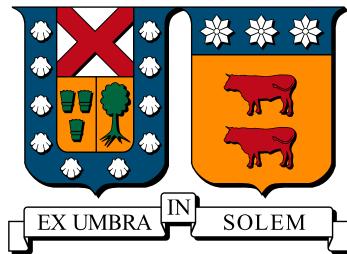
CARLOS ANDRÉS VARGAS POBLETE

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INFORMÁTICO

PROFESOR GUÍA: LIUBOV DOMBROVSKAIA

ENERO 2018

**UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
SANTIAGO – CHILE**



“TÍTULO DE LA MEMORIA”

CARLOS ANDRÉS VARGAS POBLETE

**MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INFORMÁTICO**

PROFESOR GUÍA: LIUBOV DOMBROVSKAIA

PROFESOR CORREFERENTE: PATRICIO RODRÍGUEZ

ENERO 2018

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

Agradecimientos

Resumen

El análisis de datos es un proceso fundamental para obtener información y con esta poder tomar diversas decisiones. En el ámbito escolar los establecimiento y estudiantes se agrupan por su dependencia o por la dependencia del colegio al que asisten. Este estudio busca encontrar mediante un algoritmo de clasificación no supervisado los diferentes grupos de colegios y alumnos, determinando cuales son sus principales características. Se comparan los resultados de un algoritmo al ser ejecutado con tres grupos de variables distintas y luego al mejor resultado se le agregan variables de relación establecimiento - matrículas y se compara con la primera versión. Los resultados más importantes son que los establecimientos y matrículas encuentran un óptimo de 4 grupos de clasificación con el grupo de menor cantidad de variables, donde una de las más relevantes a la hora de separar los grupos es el nivel de copago que existe.

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

Abstract

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

Índice de Contenidos

Agradecimientos	III
Resumen	IV
Abstract	V
Índice de Contenidos	VI
Lista de Tablas	VIII
Lista de Figuras	IX
Glosario	XI
Introducción	1
1. Definición del Problema	3
1.1. Datos a analizar	3
1.2. Objetivos Generales	4
1.2.1. Objetivos Específicos	4
2. Estado del Arte	5
2.1. Machine Learning	8

2.1.1. Aprendizaje Supervisado	8
2.1.2. Aprendizaje No Supervisado	9
3. Propuesta de Solución	10
3.1. Pre-procesamiento de los datos	10
3.2. X-Means	11
3.2.1. Pseudocódigo de X-Means	11
3.2.2. Ventajas	13
4. Análisis y Resultados	14
4.1. Variables	14
4.2. Experimentación	15
Conclusiones	32
A. Anexo I	33
B. Anexo II	35
Bibliografía	37

Índice de cuadros

4.1.	Clústers de establecimientos	15
4.2.	Clústers de matrículas.	16
4.3.	Clústers de matrículas.	17
4.4.	Clústers de matrículas.	20
A.1.	Atributos seleccionados y generados para la base de datos de establecimientos.	33
B.1.	Atributos seleccionados y generados para la base de datos de matrículas. . .	35
B.2.	Resumen	36

Índice de figuras

3.1. Iteración X-Means.	13
4.1. Promedios de atributos normalizados de establecimientos de la Región Metropolitana.	17
4.2. Promedios de atributos normalizados de establecimientos (con atributos relacionales) de la Región Metropolitana.	19
4.3. Promedios de atributos normalizados de matrículas de la Región Metropolitana.	22
4.4. Promedios de atributos normalizados de matrículas (con atributos relacionales) de la Región Metropolitana.	24
4.5. Mapas de clústers de establecimientos sobre mapa GSE de la Región Metropolitana.	26
4.6. Mapas de clústers de establecimientos (con atributos relacionales) sobre mapa GSE de la Región Metropolitana.	27
4.7. Mapas de calor de matrículas en clústers de establecimientos sobre mapa GSE de la Región Metropolitana.	28
4.8. Mapas de calor de matrículas (con atributos relacionales) en clústers de establecimientos sobre mapa GSE de la Región Metropolitana.	29
4.9. Mapas de calor de clústers de matrículas sobre mapa GSE de la Región Metropolitana.	30

4.10. Mapas de calor de clústers de matrículas (con atributos relacionales) sobre mapa GSE de la Región Metropolitana.	31
--	----

Glosario

CIAE: Centro de Investigación Avanzada en Educación.

GSE: Grupo Socioeconómico.

IDE: Índice de Desarrollo de la Educación.

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 [12]. 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.

Debido a esto es que con este estudio se busca encontrar, a partir de diversas variables, la estructura que tienen los establecimientos en Chile y sus estudiantes, de manera de poder realizar una mejor clasificación. Para esto se tomaron diferentes fuentes de información, las cuales fueron corregidas y estandarizadas para poder trabajar de manera sencilla con ellas.

Primero se escogieron las variables de interés pertenecientes a cada establecimiento o estudiante (sin considerar las variables que los relacionan), para luego ver cuales eran realmente relevantes seleccionar para realizar, todo esto mediante un algoritmo de aprendizaje no supervisado con el fin de no tener una categorización previa de los datos y poder deducir una clasificación directamente de los datos.

Una vez obtenidos los resultados de la prueba anterior, se le agregan las variables de relación establecimiento - matrículas y se comparan, 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 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 los diferentes tipos de aprendizaje que existen. El tercer capítulo se enfoca en la propuesta de solución planteada para el problema en estudio, destacando el porque y como se utilizó el algoritmo escogido. 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

Este trabajo realiza un estudio sobre algoritmos de clasificación no supervisados, específicamente X-Means, en bases de datos de establecimientos y estudiantes de la región metropolitana para entender cual es la estructura subyacente que presentan dichos datos. Con esto se busca poder establecer una manera de clasificar tanto a los colegios como a los alumnos sin contar con una idea preconcebida, sino que permitiendo que los datos analizados entreguen dicha información.

A la fecha no existen otros estudios que realicen un trabajo similar al planteado en el presente documento, ya que ha diferencia de este buscan establecer a que establecimiento debe asistir cada alumno según diferentes funciones de utilidad.

1.1. Datos a analizar

Las bases de datos de establecimientos y matrículas fueron facilitadas por el Centro de Investigación Avanzada en Educación (CIAE). Donde la primera fue complementada con la información disponible en la página web del Ministerio de Educación de Chile (MIME[3]) mediante la técnica de *web scraping*, pudiendo así recolectar la información pública disponible para cada colegio.

La base de datos de establecimientos corresponde a los colegios de la Región Metropolitana que imparten enseñanza básica y media para niños y jóvenes, tanto humanista científico como técnico profesional. Por otro lado las matrículas corresponden a estudiantes que se encuentran inscritos en los colegios anteriormente descritos.

1.2. Objetivos Generales

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.

Para esto se utilizará un algoritmo de clusterización sobre establecimientos y matrículas, los cuales se asociarán entre sí para determinar los atributos que los relacionan.

1.2.1. Objetivos Específicos

1. Clasificar los establecimientos mediante el uso del algoritmo de clusterización X-Means y comparar los resultados según el número de atributos escogidos para cada prueba.
2. Clasificar las matrículas mediante el uso del algoritmo de clusterización X-Means y comparar los resultados según el número de atributos escogidos para cada prueba.

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 qué variables son relevantes de considerar.

En el transcurso de los años han habido varios investigadores interesados en conocer los diferentes factores que influyen en los padres al momento de elegir un colegio en Chile.

En el 2002 Sapelli y Torche [11] estudiaron los diferentes determinantes que inciden en la elección del tipo de colegio que realizan los padres al momento de matricular a sus hijos en un determinado establecimiento educacional. En este suponen que una mayor 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 (CA-SEN) de 1996 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 resultados que obtuvieron apuntan a que algunos de los factores más determinantes son el nivel de ingreso, la educación de los padres, la recepción de subsidios y la calidad del colegio. Además destacan que por ser los subsidios por colegios y no por alumno, es decir no son portables, genera que sea más difícil para las familias de menores recursos acceder a estos si deciden optar por un colegio donde el nivel de subsidio es menor. 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 [5] buscando resolver la interrogante de cómo los padres escogen el colegio para sus hijos usaron un modelo basado en el desarrollado por McFadden [7], junto con las especificaciones planteadas por Berry, Levinsohn y Pakes [1]. Para el estudio se consideraron diferentes variables, las 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. A partir del modelo y los datos utilizados se obtuvo que existen dos variables que afectan más al momento de escoger un colegio, las cuales son el resultado del establecimiento en las pruebas y la distancia entre el hogar y el colegio, en donde la primera variable se repite respecto al estudio [11].

Dos años después, en el 2011, Daniel Gómez en conjunto con R. Chumacero y R. Paredes [4] realizan un estudio similar a los ya presentados, en donde consideran 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 [11], 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 considerados datos de la encuesta CASEN y del SIMCE, ambos correspondientes al año 2003. Mediante los estudios

¹98,2 % de cobertura educacional para el nivel de enseñanza básica en el año 1996. Fuente: CASEN 1996.

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 [6], 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 [11]. 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 [2], 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 los trabajos anteriormente señalados es importante conocer algunos conceptos de metodología que serán clave para el estudio.

2.1. Machine Learning

El *Machine Learning* o Aprendizaje automático el departamento de informática de la Universidad Técnica Federico Santa María (UTFSM) lo define como ”una subrama de la Inteligencia Artificial (IA), y como su nombre lo indica, esta tecnología trata de darle a la máquina la capacidad de aprender. El aprendizaje automático se basa algoritmos que aprenden y realizan predicciones. Tales algoritmos operan mediante la construcción de un modelo basados en conjuntos de datos de entrenamiento” [9].

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 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 es como 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, el aprendizaje supervisado y el aprendizaje no supervisado.

2.1.1. Aprendizaje 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.

2.1.2. Aprendizaje No Supervisado

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 y la extensión propuesta en [10], X-Means.

Capítulo 3

Propuesta de Solución

3.1. Pre-procesamiento de los datos

Antes de ejecutar el algoritmo de clusterización se realiza un proceso ETL (extract, transform and load), el cual permite limpiar y estandarizar las bases de datos.

Se seleccionan atributos numéricos y categóricos para los establecimientos y se generan otros a partir de la información extraída del MIME[3], debido a que mucha de esta no se encuentra estandarizada. De la misma forma se seleccionan atributos numéricos y categóricos para las matrículas, y a partir de variables no seleccionadas se calculan nuevas como es el caso de la sobre edad. Además se incorpora de manera relacional con el establecimiento al que asisten su nivel de copago y la distancia a la cual viven del colegio. De igual manera, para los establecimientos se agregan atributos relacionales, como lo es la distancia a la que viven sus estudiantes en su percentil 75 y el índice de desarrollo de la educación (IDE) por rango. El detalle de los atributos seleccionados pueden ser encontrados en los anexos A.1 y B.1 respectivamente.

Luego de esto se transforman las variables categóricas o nominales a numéricas para poder utilizarlas en el algoritmo de clusterización junto al resto de las variables numéricas.

Se imputaron los datos faltantes dentro de cada uno de los atributos previamente escogidos

mediante el algoritmo MICE (*Multiple imputation by chained equations*) para eliminar todos los valores nulos.

3.2. X-Means

X-Means es un algoritmo de agrupamiento, extendido de K-Means, propuesto por Pelleg y Moore [10] en el cual se busca dar solución a los principales problemas de K-Means. Estos son: baja escalabilidad computacional, requiere el ingreso de un número determinado de clústers y es sensible 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 de este rango el algoritmo es capaz de determinar cual es el número de centroides correcto basándose en una heurística.

3.2.1. Pseudocódigo de X-Means

El algoritmo 1 generado por Montresor y Guerrieri [8] 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 1 X-Means (simplificado) [8]

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

Require: Función 2Means(S) retorna 2 clústers

```

1: Clustering ← 2Means(S)
2: Mejor_Puntuacion ←  $-\infty$ 
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 3.1, donde se encuentra una representación para una iteración del algoritmo, con un conjunto inicial de 3 clústers.

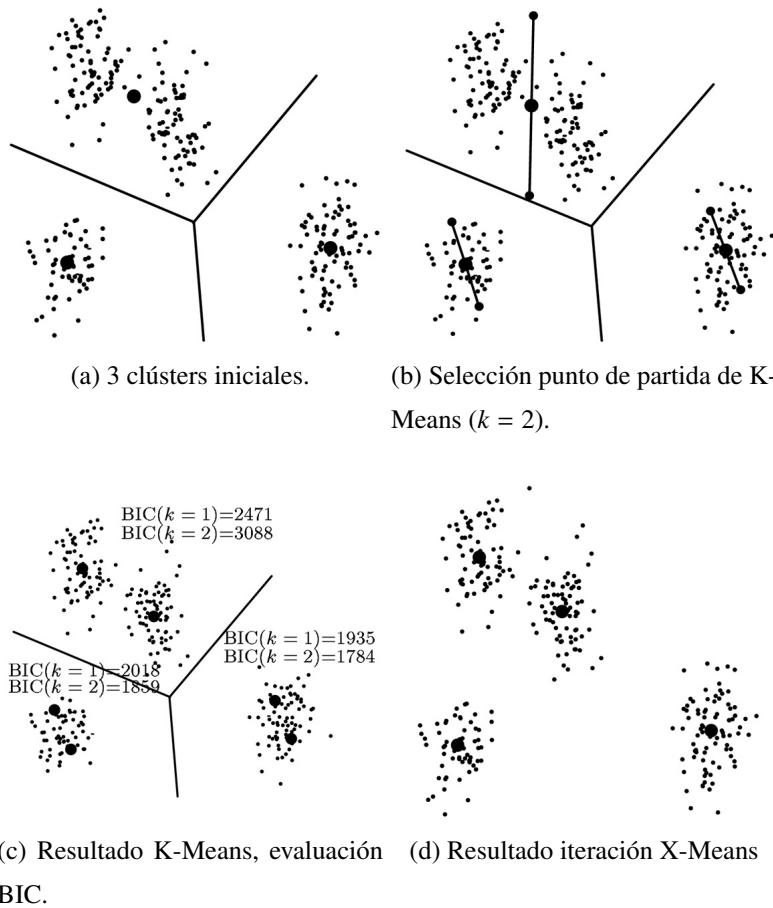


Figura 3.1: Iteración X-Means.

3.2.2. Ventajas

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.

Capítulo 4

Análisis y Resultados

En este capítulo se describe el proceso de selección de variables a considerar en el estudio, con las cuales se ejecutará el algoritmo X-Means. Luego se presentan los resultados obtenidos y sus respectivos análisis.

4.1. Variables

Las variables seleccionadas tanto para establecimientos como para las matrículas, que se encuentran individualizadas en los anexos A.1 y B.1 respectivamente, fueron filtradas según los porcentajes que ocupan cada uno de sus valores. Es decir, se contabilizó la cantidad de repeticiones para un valor dentro de la variable y se calculó su porcentaje. Con esto de clasificaron en 3 categorías según su importancia:

- Baja: cuando el porcentaje de aparición de un valor es mayor o igual a 95 %.
- Media: cuando el porcentaje de aparición de un valor es mayor o igual a 85 % y menor que 95 %.
- Alta: cuando ni uno de los valores de una variable alcanza un porcentaje de aparición mayor o igual a 85 %.

Además de esta categorización se distinguieron las variables propias del establecimiento-/matrícula y las que las relacionan. Las de relación para los establecimientos son IDE por rangos, distancia al establecimiento y nivel de sobre edad de sus alumnos. En el caso de las matrículas son distancia al colegio y nivel de copago.

4.2. Experimentación

En primera instancia se ejecutó X-Means para establecimientos y matrículas en 3 diferentes versiones, una con todas las variables, una con las de importancia alta y media, y finalmente una solo con las de alta (todas estas sin considerar las variables que los relacionan). Por tratarse de un aprendizaje no supervisado es difícil establecer una medida de eficiencia y se tomaron como válidas las siguientes versiones. En el caso de los establecimientos se puede notar en la tabla 4.1 que se mantiene constante un clúster de gran tamaño que agrupa casi un 50 % del total de colegios, en donde además en la segunda y tercera versión el resto de los clústers tiene una cardinalidad similar. Considerando lo anterior y el hecho de que al ejecutar el algoritmo con menos variables el tiempo de ejecución es menor, se tomó para estudio la tercera versión que solo considera las variables de alta importancia. Además cabe destacar que para las 3 pruebas realizadas el número óptimo de clústers es 4.

Cuadro 4.1: Clústers de establecimientos.

Variables	E_TODOS_0	E_TODOS_1	E_TODOS_2	E_TODOS_3
Todas	959	826	247	36
Alta + Media	959	540	311	258
Alta	977	524	308	259

En el caso de las matrículas, como se puede apreciar en la tabla 4.2, los resultados entre las diferentes pruebas son bastante diferentes a nivel de cardinalidad de sus clústers. Por lo tanto, considerando que la tercera prueba genera grupos de similar tamaño y que por tener una menor cantidad de variables se ejecuta en menos tiempo, se utilizó esta versión en el resto del estudio. De igual manera que para los clústers de establecimientos, para las matrículas el

óptimo de clústers es de 4.

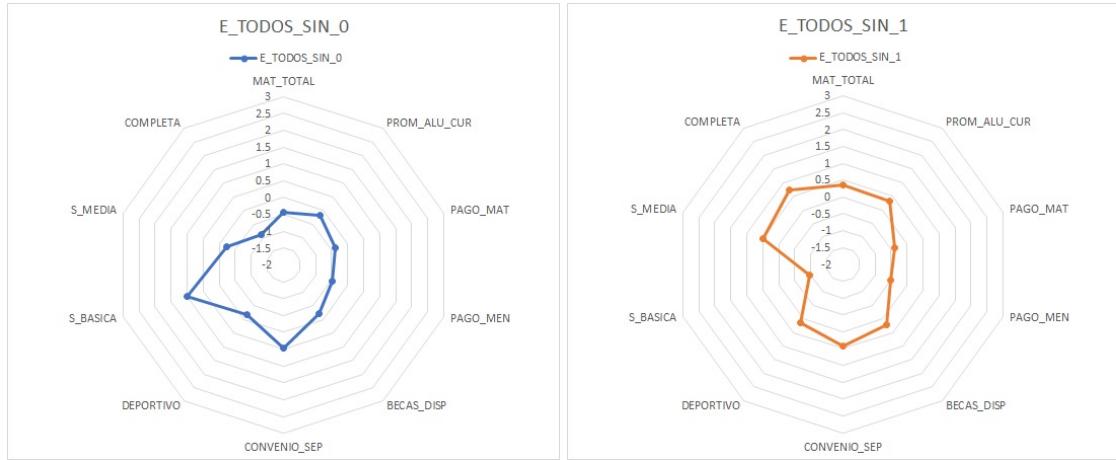
Cuadro 4.2: Clústers de matrículas.

Variables	E_TODOS_0	E_TODOS_1	E_TODOS_2	E_TODOS_3
Todas	551932	442048	13585	40803
Alta + Media	442048	551932	34088	20300
Alta	286089	300007	230486	231786

A partir de las decisiones anteriores se analizan las características de cada clúster y se repite el experimento incluyendo las variables de relación establecimiento - matrículas.

Como se puede apreciar en la figura 4.1 y la tabla 4.3 las principales características de los clústers de establecimientos son:

- E_TODOS_SIN_0: Colegios particulares subvencionados y municipales principalmente de educación básica y media. Tanto sus matrículas como sus mensualidades son gratuitas o bajas (menor a \$25.000). Poseen convenio SEP, promedio de matrículas: 383, de alumnos por curso: 26 y de becas: 19
- E_TODOS_SIN_1: Colegios principalmente particulares subvencionados, municipales y todos los de corporación de administración delegada, ofrecen educación media o completa con matrículas y mensualidades gratuitas o bajas (menor a \$25.000). Poseen convenio SEP, promedio de matrículas: 781, de alumnos por curso: 30 y de becas: 59
- E_TODOS_SIN_2: Colegios particulares subvencionados de educación media o completa con matrículas bajas (menor a \$10.000) y mensualidades medias (\$25.000 - \$100.000). En la mayoría no poseen convenio SEP, promedio de matrículas: 905, de alumnos por curso: 32 y de becas: 91.
- E_TODOS_SIN_3: Colegios particulares pagados principalmente de educación completa con matrículas y mensualidades elevadas (sobre \$50.000). No poseen convenio SEP, promedio de matrículas: 686, de alumnos por curso: 21 y de becas: 8.



(a) Clúster de establecimientos E_TODOS_SIN_0. (b) Clúster de establecimientos E_TODOS_SIN_1.



(c) Clúster de establecimientos E_TODOS_SIN_2. (d) Clúster de establecimientos E_TODOS_SIN_3.

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

Cuadro 4.3: Clústers de matrículas.

Clúster	Municipal	P. Subvencionado	P. Pagado	Corp. Admin. Del.
E_TODOS_SIN_0	485	491	1	0
E_TODOS_SIN_1	173	318	0	33
E_TODOS_SIN_2	3	303	2	0
E_TODOS_SIN_3	0	1	258	0

De la misma forma se caracterizan los clústers generados al incorporar las variables de relación establecimiento - matrículas, lo que se observa en la figura 4.2 y la tabla 4.4.

- E_TODOS_CON_0: Colegios particulares subvencionados y municipales principalmente de educación básica de matrícula gratuita y mensualidad gratuita o de precio bajo (menor a \$25.000). Poseen convenio SEP e IDE entre 0,5 y 1. Promedio de matrículas: 383, de alumnos por curso: 26 y de becas: 19. La distancia que deben recorrer sus alumnos es de 2.960 metros (percentil 75). El promedio de sobre edad es de 0,524 y 31 % son mayor o igual a 1 año. Estos se distribuyen de la siguiente forma según los años de sobre edad: 74,8 % (1), 18,9 % (2), 5,2 % (3) y 1,1 % (4).
- E_TODOS_CON_1: Colegios principalmente particulares subvencionados, municipales y todos los de corporación de administración delegada con educación media o completa, de matrículas y mensualidades gratuitas o de precio bajo (menor a \$25.000). Poseen convenio SEP e IDE entre -0,5 y 0. Promedio de matrículas: 781, de alumnos por curso: 30 y de becas: 58. La distancia que deben recorrer sus alumnos es de 5.204 metros (percentil 75). El promedio de sobre edad es de 0,488 y 32,9 % son mayor o igual a 1 año. Estos se distribuyen de la siguiente forma según los años de sobre edad: 75,1 % (1), 20,6 % (2), 3,9 % (3) y 0,5 % (4).
- E_TODOS_CON_2: Colegios particulares subvencionados de educación media o completa con matrículas bajas (menor a \$10.000) y mensualidades medias (\$25.000 - \$100.000). En su mayoría no poseen convenio SEP, nivel deportivo bajo e IDE entre 0 y 1,5. Promedio de matrículas: 895, de alumnos por curso: 32 y de becas: 91. La distancia que deben recorrer sus alumnos es de 5.207 metros (percentil 75). El promedio de sobre edad es de 0,297 y 24 % son mayor o igual a 1 año. Estos se distribuyen de la siguiente forma según los años de sobre edad: 85,3 % (1), 13 % (2), 1,6 % (3) y 0,1 % (4).
- E_TODOS_CON_3: Colegios particulares pagados principalmente de educación completa con matrículas y mensualidades elevadas (sobre \$50.000). No poseen convenio SEP y su IDE esta 0,5, pero predominantemente entre 1 y 1,5. Promedio de matrículas: 686, de alumnos por curso: 21 y de becas: 8. La distancia que deben recorrer sus

alumnos es de 9.761 metros (percentil 75). El promedio de sobre edad es de 0,488 y 38,1 % son mayor o igual a 1 año. Estos se distribuyen de la siguiente forma según los años de sobre edad: 94,9 % (1), 4,6 % (2), 0,4 % (3) y 0 % (4).



(a) Clúster de establecimientos E_TODOS_CON_0. (b) Clúster de establecimientos E_TODOS_CON_1.



(c) Clúster de establecimientos E_TODOS_CON_2. (d) Clúster de establecimientos E_TODOS_CON_3.

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

Cuadro 4.4: Clústers de matrículas.

Clúster	Municipal	P. Subvencionado	P. Pagado	Corp. Admin. Del.
E_TODOS_CON_0	485	489	1	0
E_TODOS_CON_1	173	307	0	33
E_TODOS_CON_2	3	316	2	0
E_TODOS_CON_3	0	1	258	0

Al momento de comparar ambos resultados se puede ver que al incluir las variables de relación los clústers no muestran una gran variación, pero si aumenta el nivel de detalle de cada clúster. A partir de esto se aprecia que los primeros dos clústers son colegios gratuitos o baratos con un IDE bajo que se diferencian entre ellos principalmente por el nivel de educación que imparten. El tercer clúster se diferencia de los anteriores en que su nivel de copago es de un nivel medio al igual que su IDE. Finalmente el cuarto clúster tiene un nivel de copago elevado y un mejor IDE. Un factor que los diferencia a todos es la distancia que recorren sus estudiantes para llegar al establecimiento, ya que desde el primero al cuarto la distancia va en aumento. Finalmente otro punto interesante de analizar es la sobre edad, que en el caso de los colegios mas caros se concentra con un 95 % en un año de sobre edad. En el resto de los clústers el valor fluctúa entre un 75 % y 85 %, y el resto de distribuye de 2 a 4 años de sobre edad.

De manera similar y apoyándose en la figura 4.3 se caracterizan los clústers de las matrículas, los cuales quedan de la siguiente manera:

- M_TODAS_SIN_0: Matrículas de alumnas que son beneficiarias del SEP por pertenecer a Chile solidario o por puntaje de ficha de protección social menor o igual al punto de corte. Tienen una sobre edad promedio de 0,364 y 28,5 % son mayor o igual a 1 año. Estos se distribuyen de la siguiente forma según los años de sobre edad: 76,9 % (1), 18,5 % (2), 4 % (3) y 0,7 % (4).
- M_TODAS_SIN_1: Matrículas de alumnos (hombres) que son beneficiarios del SEP por pertenecer a Chile solidario o por puntaje de ficha de protección social menor o

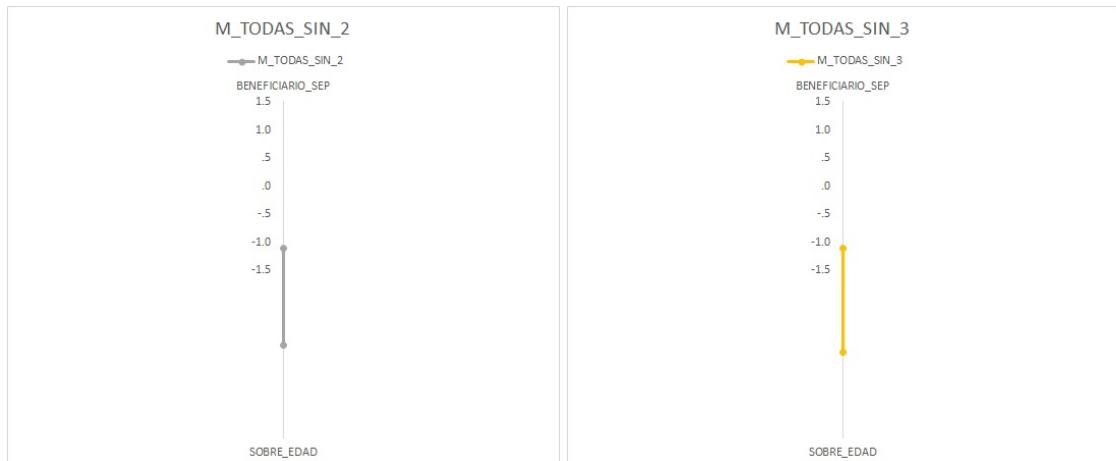
igual al punto de corte. Tienen una sobre edad de 0,488 y 36,4 % son mayor o igual a 1 año. Estos se distribuyen de la siguiente forma según los años de sobre edad: 72,5 % (1), 21,5 % (2), 5,2 % (3) y 0,9 % (4).

- M_TODAS_SIN_2: Matrículas de alumnas que no son beneficiarias SEP. Tienen una sobre edad de 0,284 y 25,7 % son mayor o igual a 1 año. Estos se distribuyen de la siguiente forma según los años de sobre edad: 90,2 % (1), 8,7 % (2), 1 % (3) y 0,1 % (4).
- M_TODAS_SIN_3: Matrículas de alumnos (hombres) que no son beneficiarios SEP. Tienen una sobre edad de 0,365 y 32 % son mayor o igual a 1 año. Estos se distribuyen de la siguiente forma según los años de sobre edad: 87, % (1), 11,1 % (2), 1,5 % (3) y 0,1 % (4).



(a) Clúster de matrículas M_TODAS_SIN_0.

(b) Clúster de matrículas M_TODAS_SIN_1.



(c) Clúster de matrículas M_TODAS_SIN_2.

(d) Clúster de matrículas M_TODAS_SIN_3.

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

Y según la figura 4.4 los clústers de matrículas incluyendo las variables de relación establecimiento - matrícula quedan caracterizados de la siguiente manera:

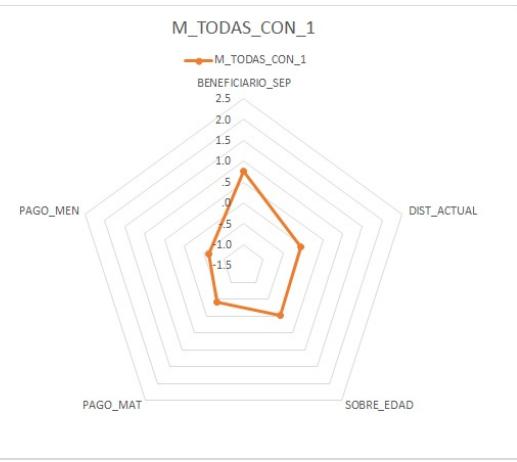
- M_TODAS_CON_0: Matrículas de alumnos (hombres) que son beneficiarios del SEP por pertenecer a Chile solidario o por puntaje ficha de protección social menor o igual al punto de corte. Asisten principalmente a colegios de matrícula y mensualidad gratuita. Tienen una sobre edad de 0,492 y 36,7 % son mayor o igual a 1 año. Estos se

distribuyen de la siguiente forma según los años de sobre edad: 72,4 % (1), 21,6 % (2), 5,1 % (3) y 0,9 % (4). La distancia que deben recorrer es de 4.939 metros (percentil 75).

- M_TODAS_CON_1: Matrículas de alumnas que son beneficiarias del SEP por pertenecer a Chile solidario o por puntaje ficha de protección social menor o igual al punto de corte. Asisten principalmente a colegios de matrícula y mensualidad gratuita. Tienen una sobre edad de 0,370 y 28,9 % son mayor o igual a 1 año. Estos se distribuyen de la siguiente forma según los años de sobre edad: 76,7 % (1), 18,7 % (2), 3,9 % (3) y 0,7 % (4). La distancia que deben recorrer es de 4.711 metros (percentil 75).
- M_TODAS_CON_2: Matrículas de alumnos (hombres y mujeres) que no son beneficiarios del SEP, que asisten principalmente a colegios con matrícula gratuita y mensualidad entre los \$10.000 y \$100.000. Tienen una sobre edad de 0,270 y 23,4 % son mayor o igual a 1 año. Estos se distribuyen de la siguiente forma según los años de sobre edad: 85,9 % (1), 12,4 % (2), 1,6 % (3) y 0,1 % (4). La distancia que deben recorrer es de 5.912 metros (percentil 75).
- M_TODAS_CON_3: Matrículas de alumnos (hombres y mujeres) que no son beneficiarios del SEP, que asisten principalmente a colegios con matrícula y mensualidad mayores a \$100.000. Tienen una sobre edad de 0,401 y 38,1 % son mayor o igual a 1 año. Estos se distribuyen de la siguiente forma según los años de sobre edad: 94,9 % (1), 4,6 % (2), 0,4 % (3) y 0 % (4). La distancia promedio que deben recorrer es de 4.911 (percentil 75).



(a) Clúster de matrículas M_TODAS_CON_0.



(b) Clúster de matrículas M_TODAS_CON_1.



(c) Clúster de matrículas M_TODAS_CON_2.



(d) Clúster de matrículas M_TODAS_CON_3.

Figura 4.4: Promedios de atributos normalizados de matrículas (con atributos relacionales) de la Región Metropolitana.

Al comparar los resultados obtenidos con y sin las variables de relación establecimiento - matrículas, se puede notar que al no utilizar dichas variables los clústers se dividen principalmente por género y por el acceso que tienen al SEP. Al incorporar las variables de copago y distancia a los establecimientos, los primeros clústers no muestran cambios significativos, siendo los últimos dos los que presentan mayores diferencias en comparación a los sin variables del relación. Dichos grupos dejan de ser excluyentes por género y el factor que

predomina es el nivel de copago que estos presentan.

Con los clústers ya realizados, y sus respectivas geolocalizaciones, se superponen sobre el mapa de grupo socioeconómico de la Región Metropolitana obteniendo los mapas de las figuras 4.5 y 4.6. Por ser tan similares los resultados de ambas pruebas (con y sin variables de relación) sus diferencias son imperceptibles en los mapas generados. En el mapa de la figura 4.5a se aprecia que los establecimientos se distribuyen uniformemente por el área metropolitana, exceptuando la zona oriente (salvo en algunos casos). Concentrándose mayoritariamente en las zonas anaranjadas, los cuales pertenecen a los deciles del 3 al 7. De la misma forma en la figura 4.5b los establecimientos se distribuyen de manera uniforme sobre las zonas anaranjadas, con la diferencia de que disminuye su cardinalidad. En la figura 4.5c los colegios se encuentran distribuidos en las zonas que comprenden los deciles del 4 al 7, pero estos se encuentran agrupados en diferentes sectores de la zona que comprenden. Finalmente el último clúster de colegios, representado en la figura 4.5d, ubica sus establecimientos en la zona oriente de la capital (deciles del 8 al 10), la cual corresponde al sector con mayor ingreso per cápita.

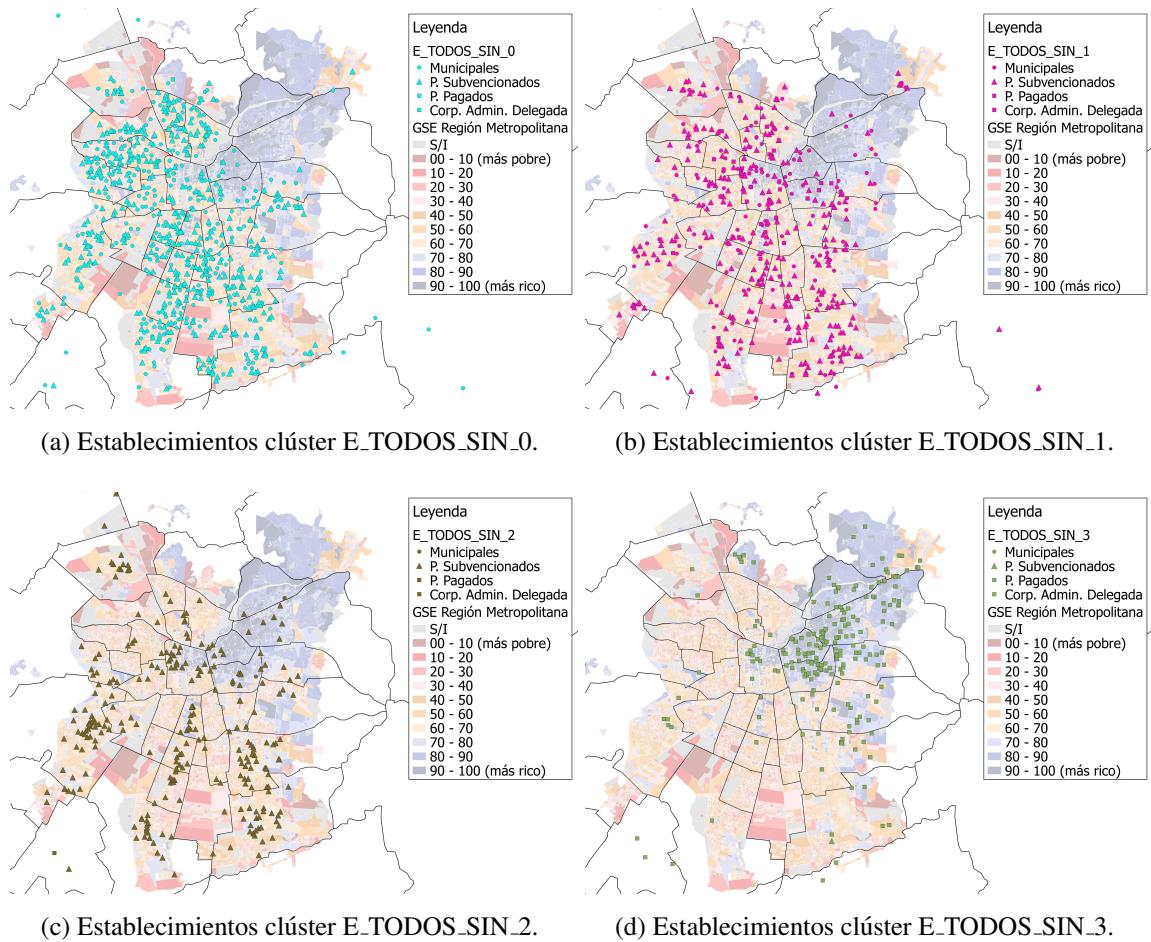
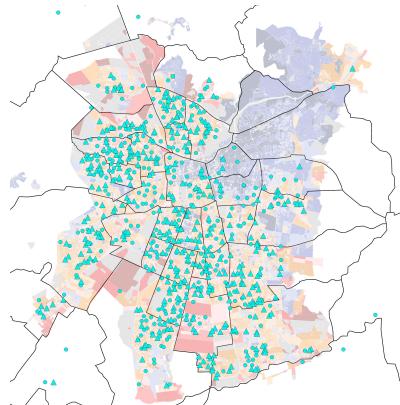


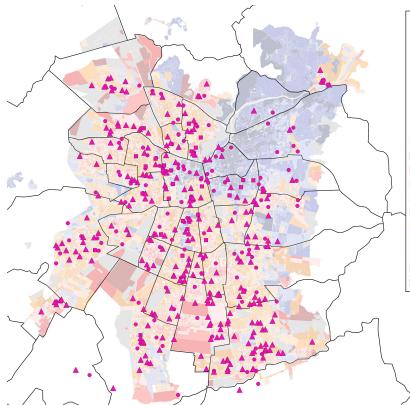
Figura 4.5: Mapas de clústers de establecimientos sobre mapa GSE de la Región Metropolitana.

Además de lo anteriormente señalado para cada clúster, se puede indicar que en su gran mayoría los establecimientos educacionales de la Región Metropolitana se encuentran en el área metropolitana y particularmente son en mayor parte del primer clúster.

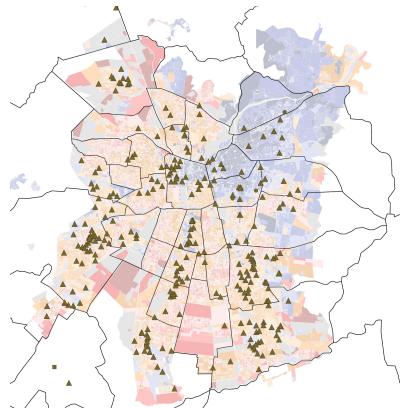
Para el caso de la figura 4.6 el análisis es similar al realizado anteriormente, debido a que el considerar o no las variables de relación establecimiento - matrícula no generan un gran impacto en la distribución de los clústers generados sin incluir dichas variables.



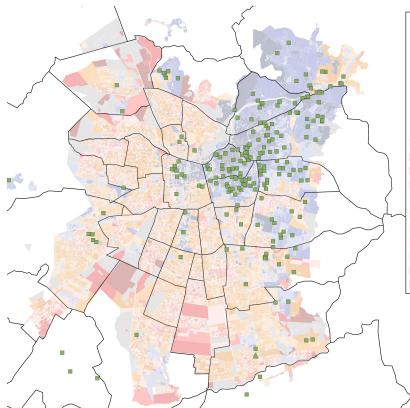
(a) Establecimientos clúster E_TODOS_CON_0.



(b) Establecimientos clúster E_TODOS_CON_1.



(c) Establecimientos clúster E_TODOS_CON_2.



(d) Establecimientos clúster E_TODOS_CON_3.



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

En las siguientes figuras (4.7 y 4.8) se muestra la distribución de los alumnos que asisten a los establecimientos de los diferentes clústers con diferentes mapas de calor. En los establecimientos del clúster E_TODOS_SIN_0 los alumnos se distribuyen por casi toda el área metropolitana, pero se concentran principalmente en un sector de la zona sur y poniente. En la siguiente (4.7b), la que representa las matrículas del clúster E_TODOS_SIN_0, los alumnos se distribuyen por toda el área, pero en menor proporción y densidad en el sector oriente. En 4.5c la distribución del alumnado se da por toda el área metropolitana, aunque se encuentra de manera mas densa en la periferia del sector poniente, seguida por dos sectores del sur de la capital. Por último, en el cuarto clúster todos sus estudiantes se sitúan en el sector oriente,

concentrándose en el sector más periférico de esta zona.

En los primeros tres clústers los alumnos que asisten a estos establecimientos viven en su gran mayoría en zonas pertenecientes a los deciles del 4 al 7, a diferencia del último caso en que sus estudiantes viven predominantemente en los sectores de deciles 8 al 10, es decir, es los lugares más acomodados del área metropolitana.

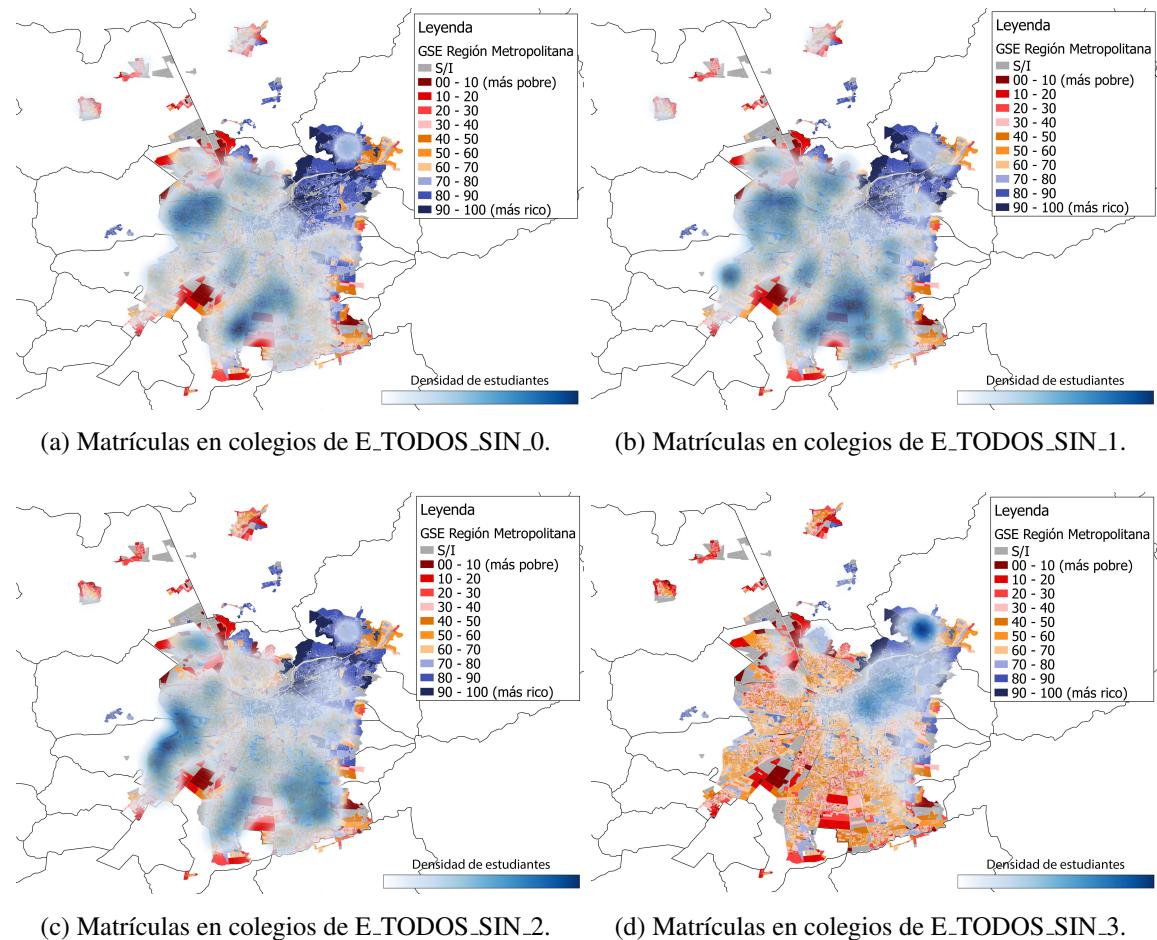


Figura 4.7: Mapas de calor de matrículas en clústers de establecimientos sobre mapa GSE de la Región Metropolitana.

Como los clúster de establecimientos no presentan cambios importantes al ser analizados con y sin variables de relación establecimiento - matrícula, sus resultados son similares, por lo que las figuras 4.7 y 4.8 son casi idénticas y es difícil percibir alguna variación entre ellas.

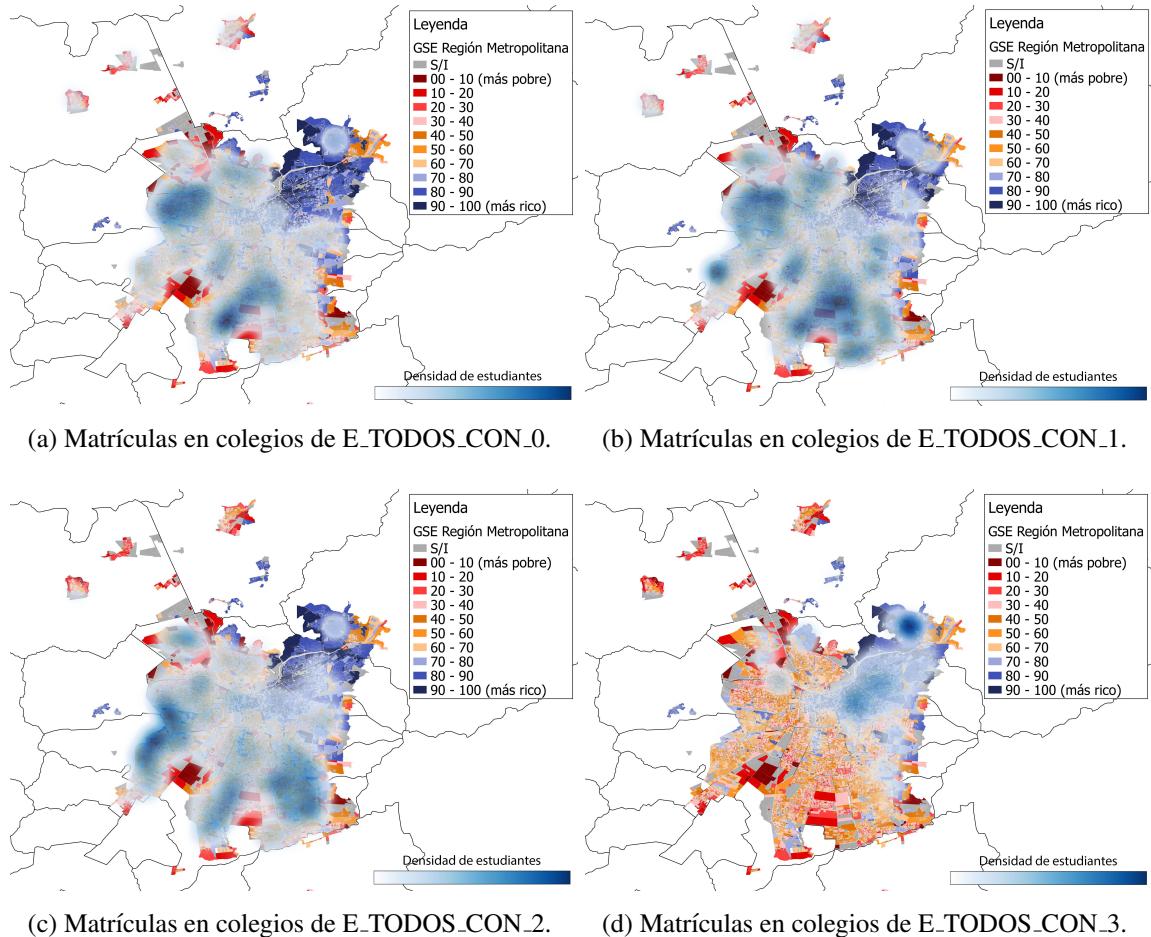


Figura 4.8: Mapas de calor de matrícululas (con atributos relacionales) en clústers de establecimientos sobre mapa GSE de la Región Metropolitana.

En las imágenes 4.9 y 4.10 se analizan las distribuciones de los clústers de estudiantes dentro del área metropolitana y sus diferentes grupos socioeconómicos, con y sin considerar sus variables de relación establecimiento - matrícula. En las figuras 4.9a y 4.9b se aprecia que los estudiantes están dispersos por casi toda el área metropolitana, en menor medida en el sector oriente y con grandes concentraciones en el sector sur y poniente. Por otro lado, en las figuras 4.9c y 4.9d si se distribuyen por toda el área metropolitana, con varios lugares de alta densidad, destacando como punto máximo la periferia del sector oriente.

Al analizar los grupos socioeconómicos en los cuales se encuentran, los primeros dos están presentes en los deciles del 4 al 7 principalmente y los otros dos en los deciles del 4 al 10,

donde su *peak* está en los deciles del 8 al 10.

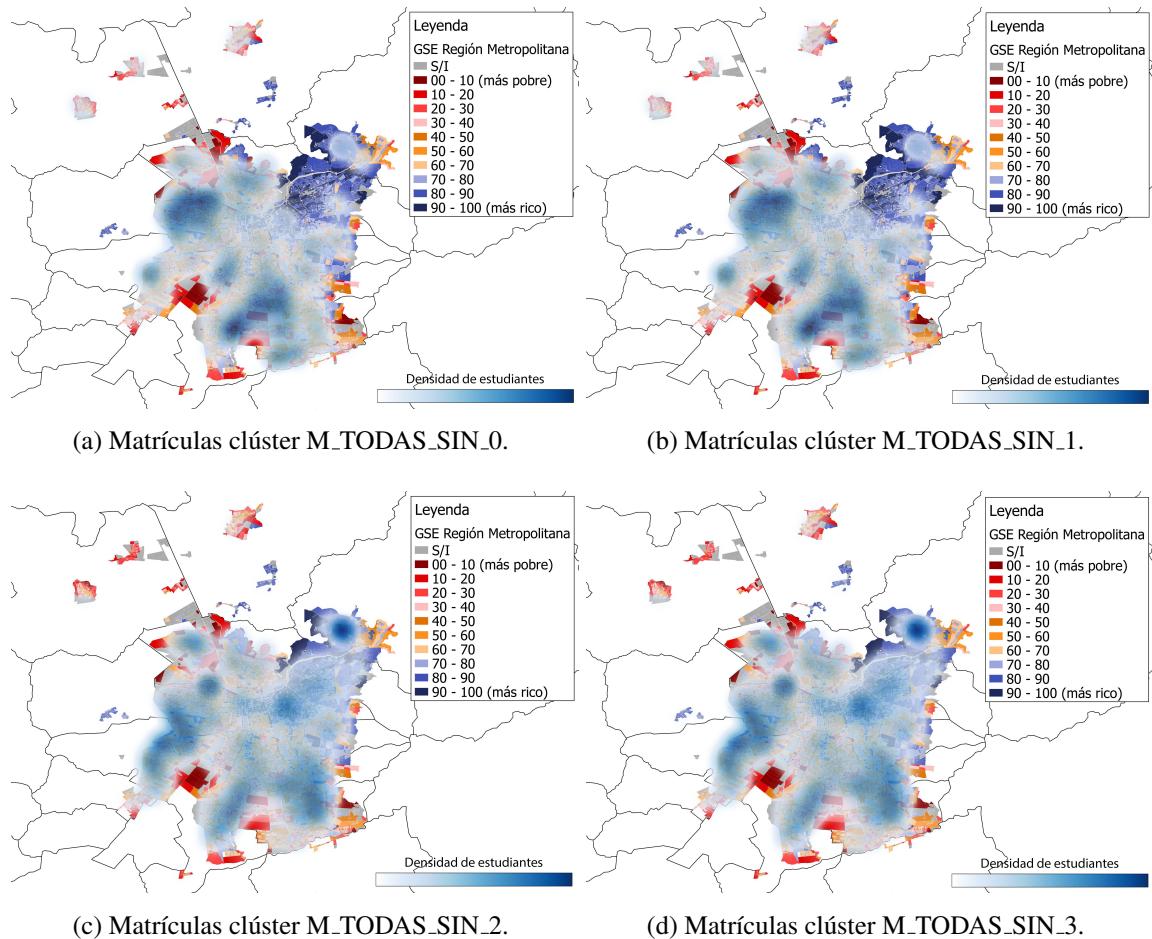


Figura 4.9: Mapas de calor de clústers de matrículas sobre mapa GSE de la Región Metropolitana.

Al analizar las imágenes 4.10a y 4.10b podemos observar una gran similitud, en donde los alumnos se distribuyen por casi toda el área metropolitana, pero disminuye notoriamente en el sector oriente. En la figura 4.10c encontramos alumnos en casi toda el área metropolitana, pero las zonas donde más se concentran los estudiantes es el sector sur y poniente. Por último, en 4.10d, casi todos los alumnos residen en el sector oriente de la capital, siendo la periferia de este donde existe una mayor concentración.

A diferencia del análisis sin variables de relación, las primeras 3 figuras muestran que los

alumnos viven en sectores socioeconómicos que van del decil 2 al decil 8, es decir, comprenden grupos muy variados. Caso aparte es la última imagen, donde los estudiantes residen en sectores pertenecientes a los deciles del 8 al 10.

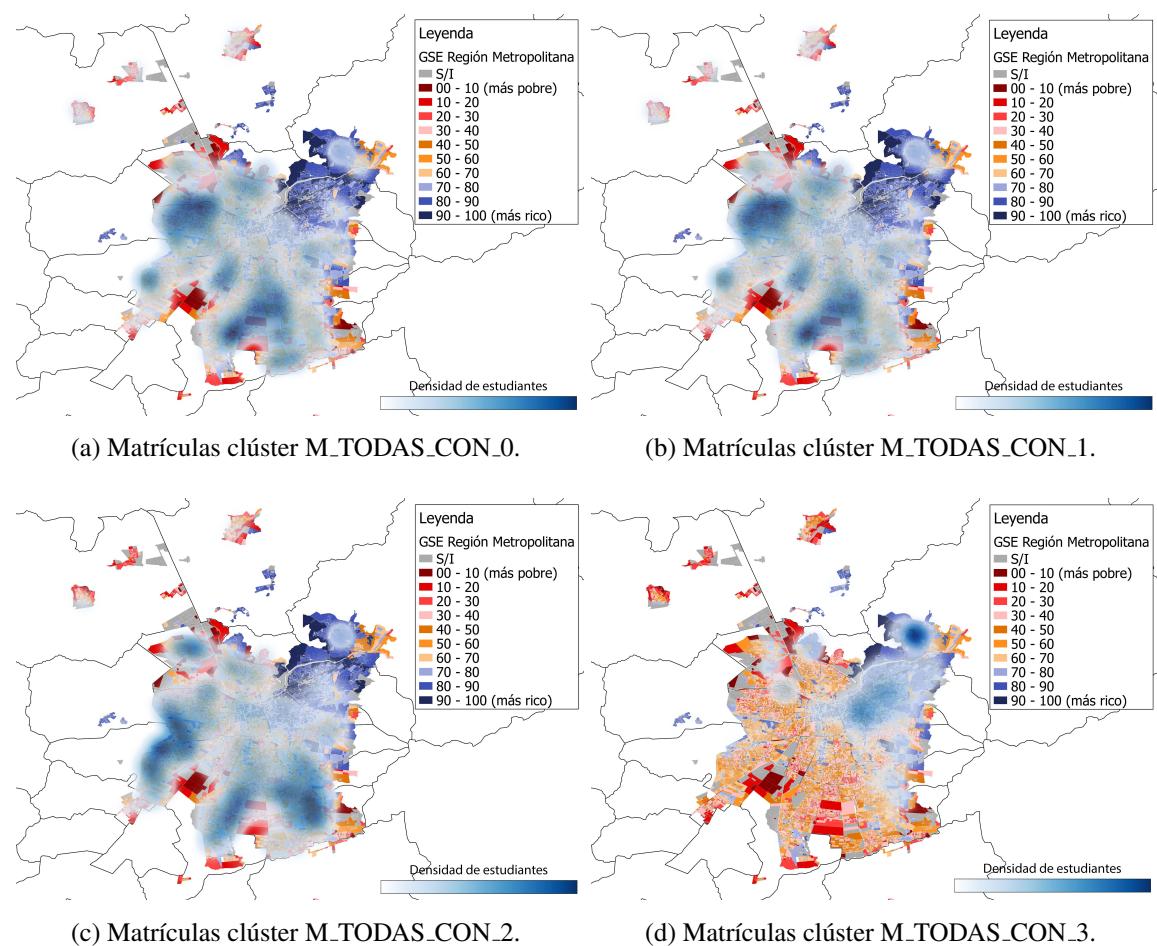


Figura 4.10: Mapas de calor de clústers de matrículas (con atributos relacionales) sobre mapa GSE de la Región Metropolitana.

Conclusiones

Anexos A

Anexo I

Cuadro A.1: Atributos seleccionados y generados para la base de datos de establecimientos.

Atributo	Descripción
area_metropolitana_rbd	Pertenencia al área metropolitana.
cod_depe	Código de dependencia del establecimiento.
gen_rbd	Género del establecimiento.
mat_total	Matrícula total de alumnos.
prom_alu_cur	Promedio de alumnos por curso.
pago_mat	Nivel de pago de matrícula.
pago_men	Nivel de pago de mensualidad.
becas_disp	Becas disponibles en el establecimiento.
convenio_sep	Posee convenio de subvención escolar preferencial (SEP).
deportivo	Nivel deportivo del establecimiento.
req_papeles	Requisitos de papeles para postular.
req_pruebas	Requisitos de prueba para postular.
req_entrevista	Requisitos de entrevista para postular.
req_pago	Requisitos de pago para postular.
continúa ...	

Cuadro A.1: Atributos seleccionados y generados para la base de datos de establecimientos.
(continuación)

req_otros	Requisitos de cualquier tipo que no clasifique en las categorías anteriores.
enf_académico	Enfoque académico.
enf_valorico	Enfoque valórico.
enf_laboral	Enfoque laboral.
enf_otros	Enfoque de otro tipo que no clasifique en las categorías anteriores.
apoyo_tutorias	Ofrece ayuda a los alumnos mediante tutorías.
apoyo_especialistas	Ofrece ayuda a los alumnos mediante especialistas.
apoyo_otros	Ofrece ayuda a los alumnos de cualquier otra forma que no clasifique en la categorías anteriores.
s_basica	Establecimiento de enseñanza básica.
s_media	Establecimiento de enseñanza media.
completa	Establecimiento de enseñanza completa.
IDE_rango	Índice de desarrollo de la educación para todos por rango.
dist_percentil_75	Distancia del percentil 75 de los alumnos que asisten al establecimiento.

Anexos B

Anexo II

Cuadro B.1: Atributos seleccionados y generados para la base de datos de matrículas.

Atributo	Descripción
area_metropolitana_alu	Pertenencia al área metropolitana.
gen_alu	Género del establecimiento.
area_metropolitana_alu	
cod_sec	Código del sector económico.
cod_espe	Código de especialidad.
cod_rama	Código de rama.
grado_sep	Corresponde a un nivel SEP.
beneficiario_sep	Indicador del alumno beneficiario de la SEP.
criterio_sep	Criterio por el cual se considera prioritario.
sobre_edad	Diferencia entre la edad actual y la esperada para el nivel.
dist_actual	Sitancia del alumno a su establecimiento actual
pago_mat	Nivel de pago de matrícula.
pago_men	Nivel de pago de mensualidad.

Cuadro B.2: Resumen

Año	CIAE	MIME	% de coincidencia
2013	2110	2040	96,68
2014	2095	2049	97,8
2015	2088	2057	98,52
2016	2068	2061	99,66

Bibliografía

- [1] Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890, 1995.
- [2] 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. *Revista Estudios Pedagógicos*.
- [3] Ministerio Chile. Mime - ministerio de educación de chile. Accedido: 7 Oct. 2017.
- [4] 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. Special Issue on Education and Health.
- [5] Francisco A Gallego and Andrés Hernando. School choice in chile: Looking at the demand side. *Pontificia Universidad Católica de Chile Documento de Trabajo*, (356), 2010.
- [6] Daniel Gómez, Rómulo A Chumacero, and Ricardo D Paredes. School choice and information. *Estudios de economía*, 39:143 – 157, 12 2012.
- [7] Daniel McFadden. Conditional logit analysis of qualitative choice. pages 105–142, 1974.
- [8] Alberto Montresor and Alessio Guerrieri. Decentralized clustering with estimation of the number of clusters. 2010.
- [9] Giovanni Eduardo Aravena Morales. Machine learning - departamento de informática. Accedido: 9 Ene. 2018.
- [10] 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.
- [11] 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.

- [12] C. Soto, X. Saavedra, I. Larraguibel, and F. Flores. Estadísticas de la educación 2016. page 15, 2017.