Universidad Nacional Autónoma de Mexico

Facultad de Ciencias

Estimación de probabilidad de gentrificación para las colonias de la CDMX: Enfoque desde la ciencia de datos

T E S I S

PARA OBTENER EL TÍTULO DE:

ACTUARIA

PRESENTA:

Yesenia Alejandro de la Cruz

TUTOR

Dr. Jorge Luis Ortega Arjona

CIUDAD UNIVERSITARIA, CD.MX.

2019

Universidad Nacional Autónoma de México
UNAM
Biblioteca Central
Dirección General de Bibliotecas de la UNAM

UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

1. Datos del alumno
Alejandro
de la Cruz
Yesenia
5571986619
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
310068804

2. Datos del tutor
Dr.
Jorge Luis
Ortega
Arjona

3. Datos del sinodal 1
Dr.
Luis Antonio
Rincón
Solís

4. Datos del sinodal 2
Dr.
León Felipe
Palafox
Novack

5. Datos del sinodal 3
M. en C.
Fernando Daniel
Pérez
Arriaga

6. Datos del sinodal 4
Act.
Edgar
Díaz
Ordoñez

7.Datos del trabajo escrito.
Estimación de probabilidad de gentrificación para las colonias de la CDMX: Enfoque desde la ciencia de datos.
94 p
2019

*"**Make no little plans**;*
They have no magic to stir men's blood and probably themselves will not be realized. Make big plans; aim high in hope and work, remembering that a noble, logical diagram once recorded will never die, but long after we are gone will be a living thing, asserting itself with ever-growing insistency. Remember that our (sons and our grandsons) are going to do things that would stagger us. Let your watchword be order and your beacon beauty.
***Think big**."*
-Daniel Burnham

Dedicado a mi madre Karina
mis hermanos Beverlyn, Samuel y Genesis
y mi novio Tamír.

II

Agradecimientos

Gracias a mi madre ***Karina*** *por darme la vida y apoyarme dentro de sus posibilidades. Gracias por tu apoyo y amor.*

Gracias a mi novio ***Tamír Vertesh*** *por el apoyo que me brindo en los retos computacionales para completar la presente tesis, por su apoyo incondicional, amor y paciencia. Gracias por ser mi equipo.*

Gracias a mi asesores ***León Felipe Palafox Novack*** *y* ***Jorge Luis Ortega Arjona*** *por la paciencia y por motivarme a explotar mi potencial en el área de Ciencia de datos.*

Resumen

La gentrificación es un fenómeno actual que se está presentando en las principales ciudades del mundo de manera paulatina. La gentrificación es un proceso de transformación donde se observa que la población de bajos ingresos es modificada y reemplazada por una población de ingresos medio y alto. Dicha población se encarga de renovar las viviendas por cuenta propia o inversión privada (agencias inmobiliarias y bancos). Todo esto impulsa la oferta, la demanda y el costo de vivienda. Como resultado, los residentes locales pueden sentirse presionados para mudarse a lugares más asequibles. En esta tesis se realiza un análisis de la gentrificación de la Ciudad de México, desde una perspectiva cuantitativa basada en datos. A partir de datos socio-demográficos, económicos y de entorno urbano se analiza el cambio de las colonias de la Ciudad de México a través del tiempo para los años 2000, 2010 y 2016. Se estudia el comportamiento de aquellas colonias que han sido gentrificadas, con el objetivo de aprender de ellas y encontrar el grupo de colonias "gentrificables", es decir, aquellas colonias que presentan una tendencia similar. Se ajusta un bosque aleatorio de clasificación para estimar la probabilidad de gentrificación para cada una de las colonias de la CDMX.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. El contexto

> "El mundo actual está experimentando un cambio social y económico, una nueva Revolución Industrial, debido en gran medida a la nueva tecnología para el procesamiento de información, representado por el crecimiento en el uso de software en la computación y las comunicaciones." (Ortega-Arjona, 2000).

Este crecimiento ha provocado que la información del mundo natural se almacene cada vez más en el ciberespacio[1] en forma de datos.

Los datos del ciberespacio muestran características de un mundo independiente, como el mundo natural (Zhu & Xiong, 2015) . Existen dos tipos de datos en el ciberespacio:

- Datos reales: Información que representa cosas del mundo natural. Un ejemplo es la información personal, datos que reflejan las características de las personas.
- Datos virtuales: Información que no representa cosas en el mundo natural. Los datos virtuales significan que las instancias de tales datos no tienen referencias en el mundo natural. Un ejemplo son las características, forma y comportamiento de una nube de puntos provenientes de un espacio n-dimensional de datos.

La formación de datos en el ciberespacio ha producido nuevos objetos de estudio y nuevos problemas. Ninguno de estos problemas son abordados por las ciencias

[1] El ciberespacio es una realidad simulada que se encuentra implementada dentro de los ordenadores y de las redes digitales de todo el mundo.

naturales o sociales; deben ser estudiados por una nueva ciencia. La **ciencia de datos** es la ciencia de estudiar los datos del ciberespacio y a la ciencia de los datos en sí misma.

Por el contrario, hay aspectos de las ciencias naturales y sociales que pueden ser estudiados desde el enfoque de la ciencia de datos. Un ejemplo es la **gentrificación**. Ésta ocurre cuando en las ciudades, la presión para la vida urbana se acelera. Estas ciudades atraen a nuevos negocios, trabajadores altamente calificados y grandes corporaciones. Generalmente surge un proceso de transformación donde se observa que la población de bajos ingresos es reemplazada por una población de ingresos medio y alto. Dicha población se encarga de renovar las viviendas por cuenta propia o inversión privada (agencias inmobiliarias y bancos). Todo esto impulsa la oferta, la demanda y el costo de vivienda. Como resultado, los residentes originales pueden sentirse presionados para mudarse.

1.2. El problema

El problema que pretende abordar esta tesis es identificar las colonias de la Ciudad de México que son "gentrificables" a partir del estudio de aquellas colonias que ya han sido gentrificadas. El objetivo es realizar un análisis cuantitativo basado en datos socio-demográficos, económicos y de entorno urbano para ajustar un modelo matemático que estime la probabilidad de gentrificación.

1.3. La hipótesis

La presente tesis busca responder a la siguientes pregunta:

> *"¿Es posible calcular la probabilidad de gentrificación de las colonias de la Ciudad de México mediante el ajuste de un algoritmo de aprendizaje supervisado basado en datos socio-demográficos, económicos y de entorno urbano?"*

1.4. La aproximación

Para calcular la probabilidad de gentrificación de las colonias de la Ciudad de México, se pretende: a) analizar los datos socio-demográficos, económicos y de

entorno urbanos de las colonias de la Ciudad de México, b) identificar correlaciones, dependencias y patrones de comportamiento en los datos analizados y; c) ajustar un algoritmo de aprendizaje supervisado que los considere con la finalidad de estimar la probabilidad de gentrificación para cada colonia de la Ciudad de México.

1.5. Estructura de la tesis

La presente tesis contiene los siguientes capítulos:

- El Capítulo 2 contiene las definiciones de los conceptos y métodos que se utilizan a lo largo de la tesis, como la definición de datos, información, conocimiento, ciencia de datos, y algunos algoritmos y modelos que se ajustan a los datos y se citan a lo largo de la tesis.
- En el Capítulo 3 se presenta trabajo relacionado del estudio de la gentrificación desde un enfoque cuantitativo basado en datos. Se incluyen dos trabajos relacionados: el primero estudia el desplazamiento de vecindarios y desarrollo de viviendas de la ciudad de Nueva York (NYC); el segundo modela el desplazamiento de los vecindarios de la bahía de San Francisco(SF) y Los Angeles (LA) mediante el ajuste de un modelo de regresión logística.
- En el Capítulo 4 se define la base de datos, se seleccionan y se construyen las variables que son utilizadas como regresores/predictores. Se ajusta un bosque aleatorio de regresión para estimar los precios del metro cuadrado para los años 2000, 2010 y 2016 basados en una muestra de datos. Se ajusta un bosque aleatorio de clasificación para estimar la probabilidad de gentrificación para cada una de las colonias de la Ciudad de México. Y finalmente se analiza la precisión del modelo.
- El Capítulo 5 contiene las conclusiones del presente estudio.

Capítulo 2

Antecedentes

En este capítulo se describe los conceptos y métodos que se utilizan a lo largo de la tesis. Se comienza por los conceptos básicos necesarios para entender la definición de ciencia de datos y gentrificación.

Se describe a la ciencia de datos comenzando por la definición de dato; mencionando los campos que se emplean en ella como la estadística, la minería de datos y el aprendizaje automático. Posteriormente, se presentan las técnicas empleadas así como una introducción a la gentrificación.

2.1. Datos

El crecimiento del uso de las tecnologías en la vida diaria provoca que la información del mundo natural se almacene cada vez más en el ciberespacio, en forma de datos.

> "La web está llena de aplicaciones basadas en datos. Casi cualquier aplicación de comercio electrónico es una aplicación basada en datos. Hay una base de datos detrás de una interfaz web y un middleware que habla con otras bases de datos y servicios de datos (compañías de procesamiento de tarjetas de crédito, bancos, etc.)"(Loukides, 2010).

El almacenamiento de los datos ciberespacio no tiene valor. Un dato no dice nada sobre el porqué de las cosas, y por sí mismo tiene poca o ninguna relevancia o propósito. El reto es procesar, organizar, estructurar y dar forma a los datos con el objetivo de convertirlos en información y después en conocimiento.

> "Un dato es una representación simbólica (numérica, alfabética, algorítmica, espacial, etc.) de un atributo o variable cuantitativa o cualitativa. La información es el procesamiento y transformación de los

> datos. Es capaz de cambiar la forma en que el receptor percibe algo, es capaz de impactar sobre sus juicios de valor y comportamientos. El conocimiento se deriva de la información, así como la información se deriva de los datos. El conocimiento es una mezcla de experiencia, valores, información y 'saber hacer' que sirve como marco para la incorporación de nuevas experiencias e información, y es útil para la acción."(Davenport & Prusak, 1998).

2.2. Ciencia de datos

> "La ciencia de los datos es la teoría, el método y la tecnología para estudiar *datanature*. Tiene dos componentes principales **El primer componente es el estudio de los patrones y las reglas de los datos en sí. Su objetivo es explorar el *datanature* y cuestiones científicas relacionadas con ello.** Esto no tiene en cuenta el significado de los datos en el mundo natural. **El segundo componente es el estudio de las reglas del mundo natural reflejado en los datos, es decir, el estudio del mundo natural realizado a través del estudio de datos.** Por ejemplo, el propósito de realizar un estudio sobre datos que representan el comportamiento de una persona es estudiar el comportamiento de esa persona." (Zhu & Xiong, 2015)

La ciencia de datos es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sea estructurados o no estructurados. Es una continuación de algunos campos de análisis de datos como la estadística, la minería de datos, el aprendizaje automático y la analítica predictiva (Liu, 2015)

2.2.1. Estadística

"La estadística es un conjunto de métodos para recolectar y analizar datos" (Agresti, 1997). Es una rama de las matemáticas que estudia usos y análisis de una muestra representativa de datos y busca explicar las correlaciones y dependencias de un fenómeno. La estadística puede dividirse en dos áreas, (Weiss, 1999):

- Estadística descriptiva: Consiste en métodos para organizar y resumir información (media, desviación estándar, histogramas, boxplot, etc)
- Estadística inferencial: Consisten en métodos para extraer y medir la fiabilidad de las conclusiones sobre la población en función de la información obtenida de una muestra de la población. Estas inferencias pueden ser pruebas

de hipótesis, pronóstico de futuras observaciones, descripciones de asociaciones o modelado de relaciones entre variables. Otras técnicas de modelación incluyen análisis de varianza, series de tiempo y minería de datos.

2.2.2. Minería de datos

La minería de datos se refiere a descubrir nuevos patrones a partir de una gran cantidad de datos, centrándose en los algoritmos para extraer conocimiento útil (Silwattananusarn & Tuamsuk, 2012).

La minería de datos tiene dos objetivos principales de predicción y descripción (Silwattananusarn & Tuamsuk, 2012):

- La predicción implica el uso de algunas variables en los conjuntos de datos para predecir valores desconocidos de otras variables relevantes (por ejemplo, clasificación, regresión y detección de anomalías).
- Descripción implica encontrar patrones y tendencias comprensibles para los humanos (por ejemplo, agrupamiento, aprendizaje de reglas de asociación y resumen).

Metodología

La metodología más utilizada es la *Cross Industry Standard Process for Data Mining* (CRISP-DM). Esta metodología divide el proceso en seis fases principales. (Olson & Delen, 2008):

- Comprensión del negocio. Involucra la determinación de objetivos de negocio, evaluación de la situación actual, establecimiento de metas y desarrollo de un plan de proyecto.
- Entendimiento de los datos. Esta etapa incluye la recopilación de datos, descripción de datos, análisis exploratorio y verificación de la calidad de los datos.
- Preparación de los datos. Una vez que las fuentes de datos disponibles son identificadas, es necesario seleccionarlas, limpiarlas y transformarlas con el objetivo de prepararlas para el modelado.
- Modelado. Se refiere a la visualización de datos y ajuste de varias técnicas de modelado. División de los datos en un conjunto de entrenamiento y de prueba.

- Evaluación. Los resultados del modelado deben ser evaluados en el contexto de negocio establecido en la primera etapa.

- Despliegue. Esta fase depende de los requerimientos, pudiendo ser simple como la generación de un reporte, o tan compleja como la puesta en producción del modelo que realice estimaciones en tiempo real.

2.2.3. Aprendizaje automático

> "El aprendizaje automático es la aplicación de un algoritmo numérico que mejora su rendimiento en una tarea determinada en función de la experiencia" (Mitchell, 1997).

El aprendizaje automático es un proceso de inducción de conocimiento centrada en el análisis de datos y la complejidad computacional del problema. De forma más concreta, se trata de crear programas capaces de generalizar comportamientos a partir de una información suministrada en forma de ejemplos.
El aprendizaje automático puede tener los siguientes enfoques:

- Arboles de decisión. Son técnicas simples que permiten estimar y explicar la relación entre algunas medidas sobre una observación y su valor objetivo. Un árbol de decisión es un modelo predictivo que puede usarse para resolver problemas de regresión y clasificación, y hacen referencia a un modelo jerárquico de decisiones y sus consecuencias(Rokach & Maimon, 2015).

- Reglas de asociación. Es una función que mapea el reporte o diagnóstico de un experto a una acción y cada acción tiene asociada una probabilidad de éxito. Comúnmente se toma la decisión más natural, es decir, se toma aquella acción con mayor probabilidad de éxito (Othman & Sandholm, 2010).

- Algoritmos genéticos. Son algoritmos de búsqueda heurística adaptativos basados en las ideas evolutivas de la selección natural y la genética. Como tales, representan una explotación inteligente de una búsqueda aleatoria utilizada para resolver problemas de optimización (Dulay, 1996).

- Redes neuronales artificiales. Es un paradigma de procesamiento de información que está inspirado en la forma en que los sistemas nerviosos biológicos, como el cerebro. El elemento clave de este paradigma es la nueva estructura del sistema de procesamiento de la información. Se compone de un gran número de elementos de procesamiento altamente interconectados (neuronas) trabajando al unísono para resolver problemas específicos(Maind & Wankar, 2014).

- Máquinas de vectores de soportes. Es una técnica robusta, precisa y efectiva para el reconocimiento y la clasificación de patrones. Es esencialmente un clasificador binario, pero puede ser adaptado para manejar tareas de clasificación multiclase (Chen *et al.*, 2011).

- Redes bayesianas. La estructura de una red bayesiana representa un conjunto de relaciones de independencia condicional que se mantienen en el dominio. Aprender la estructura del modelo de red bayesiana que representa un dominio puede revelar ideas sobre su estructura causal subyacente. Además, también se puede usar para predecir cantidades que son difíciles, costosas o poco éticas de medir, como la probabilidad de cáncer de pulmón, por ejemplo, basadas en otras cantidades que son más fáciles de obtener(Margaritis, 2003).

2.3. Algoritmos y Modelos en Ciencia de Datos

En esta sección se describen los algoritmos y modelos que son usados a lo largo de la Tesis.

2.3.1. Visualización de datos con T-SNE

T-SNE (*t-Distributed Stochastic Neighbor Embedding*) es una técnica de visualización que se encarga de mapear puntos de un espacio N-dimensional (high-d) a uno de baja dimensión (low-d) (van der Maaten, 2008).

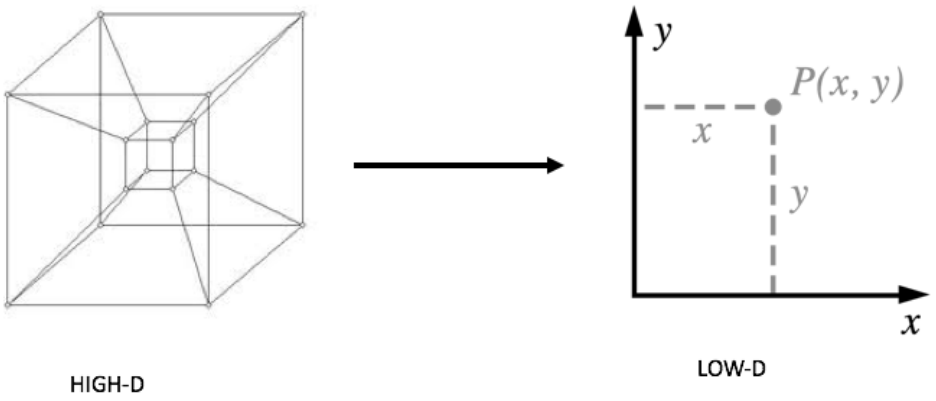

Figura 2.1: Mapeo high-d a low-d en T-SNE

T-SNE preserva la estructura del espacio N-dimensional mapeando puntos cercanos en alta dimensión a puntos cercanos en la representación de baja dimensión. Sin embargo calcular la distancia en un espacio N-dimensional y en uno de baja dimensión es muy distinta. Por lo tanto, se define una métrica en ambos espacios (high-d y low-d) que refleja la similaridad entre los puntos basados en un método probabílistico (van der Maaten, 2008).

High-d

Se define el conjunto de puntos pertenecientes al espacio N-dimensional:

$$X = \{x_1, x_2, .., x_N\}$$

Para cada punto en $x_i \epsilon X$, se toma la hipótesis de una distribución normal con centro en en x_i. Después se calcula la similaridad entre el punto x_i y $x_j \forall x_j \epsilon X$ utilizando la siguiente probabilidad condicional (van der Maaten, 2008):

$$p_{j|i} = \frac{exp(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} exp(-\left\|x_i - x_k\right\|^2 / 2\sigma_i^2)}$$

Para puntos cercanos la métrica de similaridad es alta; y para puntos lejanos la métrica es infinitesimal. Posteriormente se calcula $p_{j|i}$,para finalmente obtener:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

Esto denota la probabilidad de cercanía del punto i al punto j.

Low-d

De igual manera, se define la métrica de similaridad entre dos puntos en el espacio de baja dimensión:

$$q_{j|i} = \frac{exp(-\left\|y_i - y_j\right\|^2)}{\sum_{k \neq i} exp(-\left\|y_i - y_k\right\|^2)}$$

Si la medida de similaridad en high-d y low-d son correctamente calculadas, entonces $p_j|i$ y $q_{j|i}$ son iguale. Por lo tanto se busca minimizar la falta de coincidencia entre ellas.
Una vez definida las métricas de similaridad en el espacio High-d y Low-d, ahora se calcula la distancia entre distribuciones de probabilidad usando la métrica de divergencia KullbackLeibler de la siguiente manera (van der Maaten, 2008):

$$C = \sum_i KL(P_i \parallel Q_i) = \sum_i \sum_j p_{j|i} log \frac{p_{j|i}}{q_{j|i}}$$

donde P_i representa la distribución de probabilidad condicional sobre todos los puntos dado x_i, y Q_i representa la distribución de probabilidad condicional sobre todos los puntos en Low-d dado el punto mapeado y_j. Se busca minimizar la suma de Kullback-Leibler divergences sobre todo los puntos, usando el método de *gradient descent* usando la función de costo C definida anteriormente (van der Maaten, 2008).

La minimización de la función de costo es calculada usando el método de *gradient descent*. El gradiente tiene la siguiente forma (van der Maaten, 2008):

$$\frac{\delta C}{\delta y_i} = 2\sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j)$$

Intuitivamente, el gradiente puede ser interpretado como la fuerza resultante creada entre y_i y y_j. La cantidad entre ellas se repele o atrae dependiendo de la distancia entre estos dos puntos, si es demasiado pequeño o muy grande a partir de las similaridades entre los dos puntos en el espacio High-d.

El *gradient descent* es inicializado por un conjunto de puntos aleatorios de una distribución Gaussiana con varianza pequeña y centrada alrededor del origen, con el objetivo de que la optimización sea mas rápida, se agrega un término que de una suma exponencial decreciente que depende de los gradientes anteriores. Matemáticamente, el gradiente se actualiza de la siguiente forma (van der Maaten, 2008):

$$\gamma^{(t)} = \gamma^{(t-1)} + \eta\frac{\delta C}{\delta y_i} + \alpha(t)(\gamma^{(t-1)} - \gamma^{(t-2)})$$

En T-SNE, se usa la hipótesis de que la distribución usada en el espacio Low-d es una t-student con 1 grado de libertad (distribución Cauchy)(van der Maaten, 2008). Por lo tanto, la medida de similaridad en low-d es definida como:

$$q_{ji} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}}$$

Se decide utilizar la distribución t-student debido a que es muy similar a una normal, pero es más fácil de calcular, debido a que no contiene ningún término exponencial. Una t-student es equivalente a una mezcla infinita de normales con diferentes varianzas.

El algoritmo de manera general se ve así:

Algoritmo T-SNE
Sea **Datos** : $X = \{x_1, x_2, ..., x_n\}$ Sea los parámetros de la función de costo : perplejidad $Perp$ Sea los parámetros de optimización: número de iteraciones T, tasa de aprendizaje η, momentum $\alpha(t)$ **Resultado** : Representación de datos Low-dimensional $\gamma^{(T)} = \{y_1, y_2, ..., y_n\}$ **Inicio** : Calcular $p_{j\|i}$ con perplejidad $Perp$ Sea $p_{ij} = \frac{p_{j\|i} + p_{i\|j}}{2N}$ Solución inicial de la muestra $\gamma^{(0)} = \{y_1, y_2, ..., y_n\}$ de $\mu(0, 10^{-4}I)$ **for** $t = 1$ **to** T Calcular low-dimensional affinities q_{ij} Calcular el gradiente $\frac{\delta C}{\delta \gamma}$ Sea $\gamma^{(t)} = \gamma^{(t-1)} + \eta \frac{\delta C}{\delta \gamma} + \alpha(t)(\gamma^{(t-1)} - \gamma^{(t-2)})$ **fin**

En resumen, los parámetros usados en el algoritmo son los siguientes:

- **Perplejidad**: hipótesis de la varianza de la distribución normal en High-d. La perplejidad puede interpretarse como una medida del número efectivo de vecinos. El rendimiento de SNE es bastante robusto a los cambios en la perplejidad, y los valores típicos están entre 5 y 50.

- **tasa de aprendizaje** : velocidad de aprendizaje del algoritmo.

- **T**: # de iteraciones.

En la Figura 2.2 se muestra un ejemplo de una visualización de datos utilizando la técnica de T-SNE.

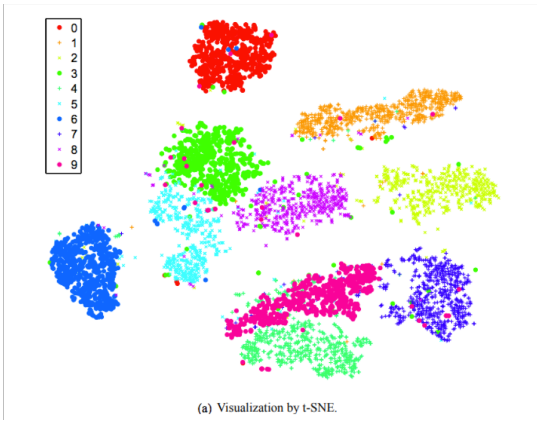

Figura 2.2: Ejemplo de visualización T-SNE

2.3.2. Arboles de decisión

Existen métodos basados en árboles para **regresión** y **clasificación**. Esto implica estratificar o segmentar el **espacio predictor** en varias regiones simples. Para hacer una predicción para una observación dada, se utiliza la media o la moda de las observaciones de entrenamiento de la región a la que pertenece. Dado que el conjunto de reglas de división utilizadas para segmentar el espacio predictor se puede resumir en un árbol, estos tipos de enfoques se conocen como métodos basados en arboles de decisión (*Decision Tree*) (Robert Tibshirani, 2013).

Arbol de Regresión

Para problemas de regresión se construye un árbol de decisión en dos pasos (Robert Tibshirani, 2013):

1. Se divide el **espacio predictor** - es decir, el conjunto de variables predictoras $X_1, X_2, ...X_p$ - en J distintas y ajenas regiones $R_1, R_2, ...R_J$,

2. Para cada observación que cae en la región R_j, se hace la misma predicción, que es simplemente la media de los valores de respuesta para las observaciones de entrenamiento en R_j

Por ejemplo, suponemos que en el paso 1 se obtiene dos regiones R_1 y R_2, y la respuesta promedio de las observaciones de entrenamiento (*training*) en la primera región es 10, mientras que la respuesta promedio en la segunda región es 20.

Entonces, dado una observación $X = x$, si $x \epsilon R_1$ se estima el valor de 10 y si $x \epsilon R_2$ se estima el valor de 20 (Robert Tibshirani, 2013).
Se divide el espacio predictor en rectángulos de alta dimensión, o cajas, para simplificar y para la fácil interpretación del modelo predictivo resultante. El objetivo es encontrar cajas $R_1, ..., R_J$ que minimizan el RSS, (*Residual Sum of Squares*) dado por (Robert Tibshirani, 2013):

$$RSS = \sum_{j=1}^{J} \sum_{i \epsilon R_j} (y_i - \widehat{y}_{R_j})^2 \tag{2.1}$$

donde $\widehat{y}_{R_j}$ es la respuesta media de las observaciones de entrenamiento dentro de la J-ésima caja. Desafortunadamente, es computacionalmente inviable considerar cada posible partición del espacio de características en J cajas. Por esta razón, se toma un enfoque abreviado, codicioso, conocido como **división binaria recursiva**. El enfoque es descendente porque comienza en la parte superior del árbol (en el que todas las observaciones pertenecen a una sola región) y luego se divide el espacio predictor; cada división se indica a través de dos nuevas ramas más abajo en el árbol. Es codicioso porque en cada paso del proceso de construcción de árboles, la mejor división se realiza en ese paso en particular, en lugar de mirar hacia adelante y escoger una división que conduzca a un mejor árbol en algún paso futuro (Robert Tibshirani, 2013).

Para realizar las divisiones binarias recursivas, primero se selecciona al predictor X_j y un punto de corte s tal que la división del espacio predictor de las dos regiones $\{X|X_j < s\}$ y $\{X|X_j \geq s\}$ minimice el RSS.[1] Se considera todos los predictores $X_1, X_2, ..., X_p$ y todos los posibles valores de puntos de corte s para cada predictor, y se escoge el predictor y el punto de corte tal que el árbol resultante tenga el mínimo RSS. Es decir, para cada j y s se define el par de semiplanos (Robert Tibshirani, 2013):

$$R_1(j, s) = \{X|X_j < s\} \quad R_2(j, s) = \{X|X_j \geq s\} \tag{2.2}$$

Y se busca el valor de j y s que minimiza la ecuación (Robert Tibshirani, 2013):

$$\sum_{i:x_i \epsilon R_1(j,s)} (y_i - \widehat{y}_{R_1})^2 + \sum_{i:x_i \epsilon R_2(j,s)} (y_i - \widehat{y}_{R_2})^2 \tag{2.3}$$

donde $\widehat{y}_{R_1}$ es la respuesta promedio de las observaciones de entrenamiento en $R_1(j, s)$ y $\widehat{y}_{R_2}$ es la respuesta promedio de las observaciones de entrenamiento en

[1] La notación $\{X|X_j < s\}$ significa la región del espacio predictor en el cual X_j tiene un valor menor que s.

$R_2(j, s)$. Se busca los valores de j y s que minimicen la ecuación 2.3 (Robert Tibshirani, 2013). Luego, se repite el proceso, buscando el mejor predictor y el mejor punto de corte para dividir los datos restantes y minimizar el RSS dentro de cada una de las regiones resultantes. Sin embargo, esta vez, en lugar de dividir todo el espacio predictor, se divide una de las dos regiones previamente identificadas. Ahora se tiene tres regiones. Nuevamente, se busca dividir una de estas tres regiones más para minimizar el RSS. El proceso continúa hasta que se alcanza un criterio de parada; por ejemplo, se puede continuar hasta que ninguna región contenga más de cinco observaciones Una vez que las regiones $R_1, ..., R_j$ se ha creado, se redirige la respuesta para una observación de prueba determinada utilizando la media de las observaciones de entrenamiento en la región a la que pertenece la observación de prueba (Robert Tibshirani, 2013).

La Figura 2.3 muestra un ejemplo de división de un espacio predictor en regiones simples.

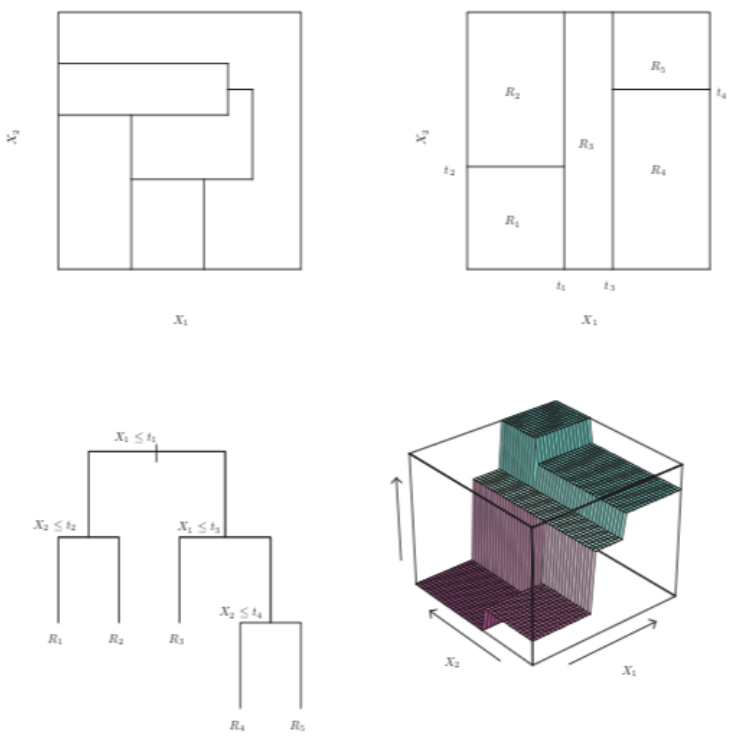

Figura 2.3: Ejemplo de espacio predictor

Arbol de Clasificación

Un árbol de clasificación es muy similar a un árbol de regresión, excepto que

se usa para predecir una respuesta cualitativa en lugar de cuantitativa. Para un árbol de regresión, la respuesta estimada para una observación viene dada por la respuesta media de las observaciones de entrenamiento que pertenecen al mismo nodo terminal. En contraste, para un árbol de clasificación, predecimos que cada observación pertenece a la clase de observaciones de entrenamiento más comunes en la región a la que pertenece. Al interpretar los resultados de un árbol de clasificación, a menudo interesa no solo la predicción de clase correspondiente a una región de nodo terminal particular, sino también las proporciones de clase entre las observaciones de entrenamiento que caen en esa región (Robert Tibshirani, 2013).

Ajustar un árbol de clasificación es bastante similar a ajustar un árbol de regresión. Al igual que en el ajuste de regresión, usamos división binaria recursiva para ajustar un árbol de clasificación. Sin embargo, en la clasificación el RSS no se puede utilizar como criterio para realizar divisiones binarias (Robert Tibshirani, 2013).

Una alternativa natural al RSS es la tasa de error de clasificación (*classification error rate*). Ya que se planea asignar una observación en una región dada a la clase más frecuente de las observaciones de entrenamiento en esa región, la tasa de error de clasificación es simplemente la fracción de las observaciones de entrenamiento en esa región que no pertenecen a la clase más común(Robert Tibshirani, 2013).

$$E = 1 - \underset{k}{max}(\widehat{p}_{mk}) \tag{2.4}$$

donde $\widehat{p}_{mk}$ representa la proporción de observaciones de entrenamiento en la m-ésima región que viene de la clase k. Sin embargo, resulta que el error de clasificación no es lo suficientemente sensible para el ajuste de árboles, y en la práctica son preferibles otras dos medidas(Robert Tibshirani, 2013).

El índice de Gini es definido por (Robert Tibshirani, 2013):

$$G = \sum_{k=1}^{K} \widehat{p}_{mk}(1 - \widehat{p}_{mk}) \tag{2.5}$$

Es una medida de la varianza total a través de las K clases. No es difícil ver que el índice de Gini toma un valor pequeño si todos los $\widehat{p}_{mk}$ están cerca de cero o uno. Por esta razón, el índice de Gini se conoce como una medida de la pureza del nodo: un valor pequeño indica que un nodo contiene predominantemente observaciones de una sola clase (Robert Tibshirani, 2013).
Una alternativa al índice de Gini es la La entropía cruzada (*cross-entropy*), dado por (Robert Tibshirani, 2013):

$$D = -\sum_{k=1}^{K} \widehat{p}_{mk} log(\widehat{p}_{mk}) \tag{2.6}$$

Entonces $0 \leq \widehat{p}_{mk} \leq 1$, si y solo sí $0 \leq -\widehat{p}_{mk} log(\widehat{p}_{mk})$. La entropía cruzada tiene un valor cercano a cero si los $\widehat{p}_{mk}$ están todos cerca de cero o cerca de uno. Por lo tanto, al igual que el índice de Gini, la entropía cruzada tiene un valor pequeño si el m-ésimo nodo es puro. De hecho, resulta que el índice de Gini y la entropía cruzada son bastante similares numéricamente (Robert Tibshirani, 2013).

Cuando se construye un árbol de clasificación, el índice de Gini o la entropía cruzada se usan normalmente para evaluar la calidad de una división en particular, ya que estos dos enfoques son más sensibles a la pureza de los nodos que la tasa de error de clasificación. Cualquiera de estos tres enfoques puede usarse al recortar el árbol, pero la tasa de error de clasificación es preferible si la meta es la precisión de la predicción del árbol recortado final(Robert Tibshirani, 2013).

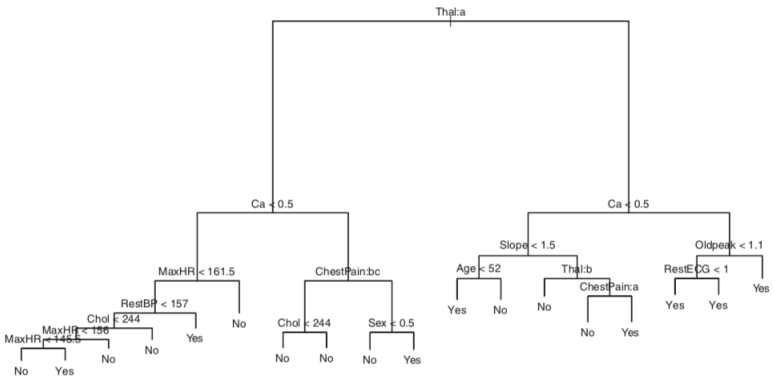

Figura 2.4: Ejemplo de árbol de clasificación

La Figura 2.4 muestra un ejemplo de árbol de clasificación.

2.3.3. *Bagging*

Boostrap agreggation o *Bagging* es un procedimiento cuyo propósito es reducir la varianza de un método de aprendizaje estadístico (Robert Tibshirani, 2013).
Sea $Z_1, Z_2, ...Z_n$ un conjunto de n observaciones independientes con varianza σ^2, entonces la varianza del promedio $\overline{Z}$ es $\frac{\sigma^2}{n}$. En otras palabras, el promedio del

conjunto de observaciones reduce la varianza (Robert Tibshirani, 2013).

Una manera de reducir la varianza e incrementar la exactitud de un modelo es tomar muchas muestras del conjunto de entrenamiento de la población, construir un modelo para cada conjunto de entrenamiento y calcular el promedio de las estimaciones. En otras palabras es posible generar B diferentes muestra *boostrap* del (único) conjunto de entrenamiento. Entrenar el modelo con el conjunto de entrenamiento b para obtener $\widehat{f}^{*b}(x)$ y finalmente obtener el promedio de todas las estimaciones para obtener (Robert Tibshirani, 2013):

$$\widehat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\widehat{f}^{*b}(x) \tag{2.7}$$

La ecuación 2.7 es llamada *bagging* (Robert Tibshirani, 2013).

2.3.4. Bosques Aleatorios

Los bosques aleatorios (*Random Forest*) proporcionan una mejora sobre el *bagging*. Al igual que en *bagging*, construimos árboles de decisión con muestras de entrenamiento *bootstrapped*. Pero cuando se construyen estos árboles de decisión, cada vez que se considera una división en un árbol, se elige una muestra aleatoria de m predictores como candidatos divididos del conjunto completo de p predictores. La división puede usar solo uno de esos m predictores. Se toma una nueva muestra de m predictores en cada división. Generalmente se elige $m \approx \sqrt{p}$, es decir, el número de predictores considerados en cada división es aproximadamente igual a la raíz cuadrada del número total de predictores (Robert Tibshirani, 2013).

En otras palabras, al construir un bosque aleatorio, en cada división del árbol, el algoritmo no considera la mayoría de los predictores disponibles. Esto tiene la siguiente razón. Supongamos que hay un predictor muy fuerte en el conjunto de datos, junto con un número de otros predictores moderadamente fuertes. Luego, en la colección de árboles *bagging*, la mayoría o todos los árboles usarán este predictor fuerte en la división superior. En consecuencia, todos los árboles *bagging* se ven bastante similares entre sí. Por lo tanto, las predicciones de los árboles *bagging* están altamente correlacionadas. Desafortunadamente, promediar muchas cantidades altamente correlacionadas no conduce a una reducción tan grande de la varianza como promediando muchas cantidades no correlacionadas. En particular, esto significa que el bagging no conduce a una reducción sustancial de la variación en un solo árbol en esta configuración. (Robert Tibshirani, 2013)

Los bosques aleatorios superan este problema obligando a cada división a considerar solo un subconjunto de los predictores. Por lo tanto, en promedio $(p-m)/p$ de las divisiones ni siquiera considera el predictor fuerte, por lo que otros predictores tienen más posibilidades (Robert Tibshirani, 2013).

La principal diferencia entre *bagging* y bosque aleatorios es la elección del tamaño del subconjunto de predictor m. Por ejemplo, si un bosque aleatorio se construye utilizando $m = p$, entonces esto equivale simplemente a *bagging* (Robert Tibshirani, 2013).

2.4. Métricas de evaluación de modelos

Para evaluar el desempeño de un método de aprendizaje estadístico en un conjunto de datos dado, se necesita alguna forma de medir qué tan bien sus predicciones realmente coinciden con los datos observados. Es decir, se requiere cuantificar la extensión a la que el valor de respuesta previsto para una observación dada está cerca de el verdadero valor de respuesta para esa observación (Robert Tibshirani, 2013). Ajustando el modelo se mide el error de ajuste y el error de predicción.

Error de ajuste o *Training Error* Se refiere al error observado para el conjunto de datos pertenecientes al conjunto de entrenamiento. Con los que se realiza el ajuste del modelo.

Error de predicción o *Test Error* Se refiere al error que se observa para un nuevo conjunto de datos con el modelo especificado, dado el ajuste realizado con los datos del conjunto de entrenamiento.

Para la evaluación del modelo se utilizan métricas. Sin embargo, depende de la naturaleza del problema. Las métricas usadas para un problema de regresión no son siempre usadas para problemas de clasificación.

2.4.1. Métricas para Regresión

Error Cuadrático Medio o *MSE: Mean Square Error*
En el ajuste de regresión, la medida más utilizada es el error cuadrático medio (MSE) (Robert Tibshirani, 2013). Intuitivamente es interpretado como el valor cuadrático promedio de error para cada observación:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \tag{2.8}$$

Error Cuadrático Medio o *MAE : Mean Absolute Error*
MAE es interpretado como el valor absoluto promedio de error para cada observación estimada (Robert Tibshirani, 2013):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i| \tag{2.9}$$

Suma de cuadrados residual o *RSS : Residual Sum of Squares*
RSS es el valor cuadrático de error del modelo y es interpretado como la cantidad de variabilidad que queda sin explicar después de realizar la regresión (Robert Tibshirani, 2013):

$$RSS = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \tag{2.10}$$

Error estándar residual o *RSE : Residual Standard Error*
RSE es una estimación de desviaciones estándar de errores. En términos generales, es la cantidad promedio que la respuesta se desvía de la verdadera línea de regresión. Se calcula utilizando la fórmula (Robert Tibshirani, 2013):

$$RSE = \sqrt{\frac{1}{n-2}RSS} \tag{2.11}$$

$\mathbf{R}^2$
La estadística R^2 proporciona una medida alternativa de ajuste. Toma la forma de una proporción, la proporción de la varianza explicada, por lo que siempre toma un valor entre 0 y 1 (Robert Tibshirani, 2013):

$$R^2 = 1 - \frac{RSS}{TSS} \tag{2.12}$$

donde $TSS = \sum_{(} y_i - \overline{y})^2$ es la suma total de cuadrados (*total sum of squares*). y RSS es definida en la ecuación 2.11. El TSS mide la varianza total en la respuesta Y, y puede considerarse como la cantidad de variabilidad inherente en la respuesta antes de que se realice la regresión. En contraste, RSS mide la cantidad de variabilidad que queda sin explicar después de realizar la regresión. Por lo tanto, TSS - RSS mide la cantidad de variabilidad en la respuesta que se explica al realizar la regresión y R^2 mide la proporción de variabilidad en Y que puede explicarse utilizando X. Una estadística R^2 que está cerca de 1 indica que

una gran proporción de la variabilidad en la respuesta ha sido explicada por la regresión. Un número cercano a 0 indica que la regresión no explica gran parte de la variabilidad en la respuesta; esto puede ocurrir porque el modelo lineal es incorrecto, o cuando el error inherente σ^2 es alto, o ambos (Robert Tibshirani, 2013).

2.4.2. Métricas para Clasificación

Las métricas usadas para la evaluación de un modelo con variable objetivo cualitativa más comunes son las siguientes.

Matriz de confusión
En la práctica, un clasificador binario puede cometer dos tipos de errores: puede asignar incorrectamente a una observación que por defecto está en la categoría no predeterminada, o puede asignar incorrectamente a una observación que no está predeterminada a la categoría predeterminada. A menudo es interesante determinar cuál de estos dos tipos de errores se están cometiendo (Robert Tibshirani, 2013). Y son mostradas en la matriz de confusión.

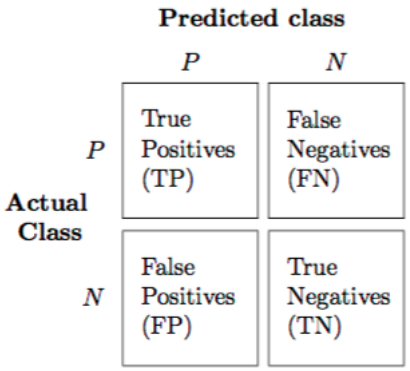

Figura 2.5: Diagrama de matriz de confusión

La matriz de confusión permite comparar el valor real de la variable objetivo (*variable target*) con el valor estimado por el modelo. Los estimaciones correctas del modelo se encuentran en a diagonal de la matriz, el valor bajo la etiqueta TN: *True Negative* o Verdaderos Negativos y TP: *True Positive* o Verdaderos Positivos. TN representa a las estimaciones cuya variable objetivo pertenece a la clase 0 y el modelo las estimo correctamente. Por el contrario TP son aquellas observaciones que pertenecen a la clase 1 y el modelo las estima de manera correcta. Un modelo perfecto solo tendría valores en esta diagonal.

Por otro lado está la diagonal inversa con los valores FP: *False Positive* o Falsos Positivos y FN : *False Negative* o Falsos Negativos.

Exactitud de la clasificación o *Clasification accuracy*
Esta métrica muestra la proporción de estimaciones correctas realizadas por el modelo.

$$Clasificationaccuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.13}$$

Sensibilidad o *Sensitivity*
El rendimiento específico de cada clase también es importante, donde los términos sensibilidad y especificidad caracterizan el rendimiento de un clasificador o prueba de detección (Robert Tibshirani, 2013).También llamada tasa de verdad positiva, recuperación o probabilidad de detección en algunos campos. Mide la proporción de positivos reales que se identifican correctamente como tales (por ejemplo, el porcentaje de personas enfermas que están correctamente identificadas como que tienen la condición).

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.14}$$

Precisión
También llamada tasa negativa verdadera. Mide la proporción de negativos reales que se identifican correctamente como tales (por ejemplo, el porcentaje de personas sanas que están correctamente identificadas como que no tienen la condición).

$$Precision = \frac{TP}{TP + FP} \tag{2.15}$$

F1 Score
Esta métrica muestra un balance entre la métrica de Precisión y Sensibilidad.

$$F1Score = 2 * \frac{Sensitivity * Precision}{Sensitivity + Precision} \tag{2.16}$$

2.5. Una breve introducción a la gentrificación

El desplazamiento poblacional es un tema importante en las ciudades y/o centros de conocimiento, donde la presión para la vida urbana se está acelerando. Estas ciudades atraen a nuevos negocios, trabajadores altamente calificados y grandes corporaciones. Generalmente en los barrios populares de estas ciudades surge un proceso de transformación donde se observa que la población de bajos ingresos es modificada y reemplazada por una población de ingresos medio y alto. Dicha población se encarga de renovar las viviendas por cuenta propia o inversión privada (agencias inmobiliarias y bancos). Todo esto impulsa la oferta, la demanda y el costo de vivienda. Como resultado, los residentes locales pueden sentirse presionados para mudarse a lugares más asequibles. Dicho proceso es definido bajo el término de ***gentrificación*** (Salinas-Arreortua, 2013).

Generalmente hay tres factores dinámicos que pueden ser identificados como importantes en la gentrificación: a) movimiento de personas,b) políticas e inversiones públicas y c) flujos de capital privado. Estos elementos no son en absoluto excluyentes -en realidad son muy dependientes entre sí- y cada uno de ellos está mediado por concepciones de clase, lugar y escala. (Salinas-Arreortua, 2013)

> "La gentrificación es típicamente el resultado de la inversión en una comunidad por parte del gobierno local, activistas comunitarios o grupos empresariales, y a menudo puede estimular el desarrollo económico, atraer negocios y reducir los índices de delincuencia" (Clark, 2004).

Diversos estudios analizan a la gentrificación. Principalmente, se centran en analizar las ventajas y desventajas desde una perspectiva social. Debido a que no existe una definición de gentrificación universalmente aceptada, es por eso que en el presente trabajo consideramos los siguientes elementos como parte integrante del proceso de gentrificación (Salinas-Arreortua, 2013):

- Movimientos de población, pudiendo presentarse desplazamiento directo o indirecto. El desplazamiento directo consiste en que la gente es forzada a dejar sus viviendas mediante acciones violentas, conocidos como desalojos compulsivos. El desplazamiento indirecto serían a partir de razones socioeconómicas; los viejos residentes son obligados a dejar sus viviendas porque aumentan los alquileres o porque aumentaron los impuestos inmobiliarios cuando el valor de mercado de la propiedad aumenta. O bien, porque las transformaciones han generado que los viejos residentes ya no se sienten cómodos de estar en su barrio. Estos desplazamientos directos e indirectos

por nuevos residentes generan mayor presencia de grupos sociales con mayores recursos económicos, con estudios profesionales, artistas, intelectuales e incluso de colectivos con rasgos en común (estudiantes, turistas, etc).

- Las viviendas degradadas o no degradadas suelen ser rehabilitadas, renovadas o cambian su uso de suelo.

- Negocios comerciales como restaurantes, estéticas, galerías de arte y bares se establecen en espacios antes ocupados por un comercio tradicional y/o familiar.

- Instalaciones antiguas (almacenes, fábricas, estaciones ferroviarias, puertos comerciales) son reconvertidos en instalaciones de diverso uso (comercial, habitacional, oficinas y servicios) destinadas a grupos sociales de ingresos medios y altos.

2.6. Colonias gentrificadas en CDMX

La gentrificación en la CDMX ya ha sido citada en algunos periódicos anteriormente. Tal es el caso de (Cantera, 2017), donde cita a la colonias **Doctores, Obrera, Tabacalera y Álamos** como colonias que atraviesan por un proceso de gentrificación debido a su oferta comercial, oferta de servicios y conectividad.

Por otro lado (CAMHAJI, 2017) cita a las colonias **Condesa , Cuauhtémoc, Hipódromo de la Condesa, Roma Sur, Roma Norte, Juárez, Doctores, Guerrero, Santa María La Ribera y Centro** en proceso de gentrificación debido al aumento de cuotas anuales de predial, aumento en los alquileres y precios de venta de inmuebles.

Según Forbes (Forbes, 2014), las colonias que tienen potencial por su posición geográfica, servicios de transporte, ofertas de entretenimiento y gastronómicas son **Juárez, Doctores, San Rafael, Santa María La Ribera, Irrigación y Escandón**.

2.7. Concentración de extranjeros en CDMX

En la Figura 2.6 se observa la concentración de extranjeros por delegación y se observa que en su mayoría están concentrados en las delegaciones Miguel Hidalgo, Benito Juárez y Cuauhtémoc.

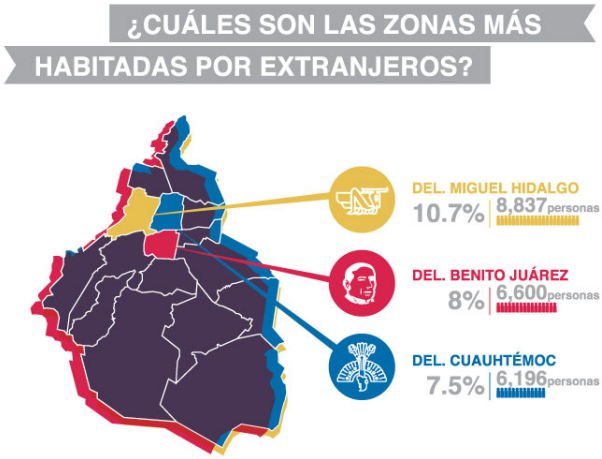

Figura 2.6: Mapa de concentración de extranjeros por delegación en 2012 (MXCity, 2019)

Por otro lado, en 2017 el Universal publicó estadísticas de colonias donde se concentran los extranjeros y cito a la comunidad estadounidense, europea y asiática (Riquelme, 2017).

> "Los estadounidenses prefieren vivir en colonias como San Ángel, al sur de la ciudad; en la colonia Juárez, en el centro, y en la zona del corredor de Reforma y Chapultepec, debido a la cercanía de la embajada de Estados Unidos en México".
>
> "Otra comunidad con una tradición y una participación relevante dentro de la comunidad de la Ciudad de México son los españoles, quienes se afincaron en colonias del norte de la ciudad como Lindavista. San Ángel, Coyoacán, la Condesa y la colonia Roma son los lugares preferidos por los europeos para vivir"
>
> "En este sentido, contrasta el caso de las personas de origen asiático que vienen a asentarse en la Ciudad de México, quienes prefieren la colonia las Águilas y la periferia de la ciudad, o bien, permanecer cerca de la avenida de los Insurgentes y sus alrededores."

2.8. Resumen

Este capítulo contiene los algoritmos y modelos que son usados a lo largo de la tesis, también se presenta las métricas que miden la exactitud de los modelos.

La gentrificación se ha estudiado desde hace varios años, sin embargo en pocas ocasiones ha sido estudiada desde un enfoque basado en datos. Por lo tanto, en la presente tesis se analiza la gentrificación desde dicha perspectiva.

Se han reportado casos de colonias en proceso de gentrificación citadas en periódicos las cuales se ubican en las delegaciones Cuauhtémoc, Benito Juárez y Miguel Hidalgo.

Capítulo 3

Trabajo relacionado

La gentrificación es un proceso social que varias disciplinas han estudiado usando sus propios métodos multidisciplinarios. En este capítulo se analizan dos trabajos de investigación relacionados con el análisis de la gentrificación desde un enfoque cuantitativo basado en datos:

- Proyecto mapa de alerta de desplazamiento de vecindarios y desarrollo de vivienda (ANHD, 2015).
- Proyecto de desplazamiento urbano (UCB & UCLA, 2017).

3.1. Proyecto mapa de alerta de desplazamiento de vecindarios y desarrollo de vivienda

La Asociación para el Desarrollo de Vecindarios y Viviendas (Association for Neighborhood and Housing Development, ANHD) ha desarrollado el Mapa de Proyecto de Alerta de Desplazamiento (Displacement Alert Project Map o DAP Map), que evalúa el riesgo de desplazamiento edificio por edificio en la ciudad de Nueva York (ANHD, 2015).

El análisis utiliza datos públicos de 96,000 edificios de Nueva York y asigna a cada uno una medida (o *score*) de riesgo basado en tres factores (ANHD, 2015):
a) pérdida de unidades reguladas por renta,
b) tasa de rotación del inquilino y
c) precio de venta del edificio.

Estandariza cada uno de los tres factores en una escala de de 0 a 100, los agrega a la medida final con el objetivo de obtener una medida de de riesgo combinado que

va en una escala de 0 a 300, y despliega el resultado en un mapa interactivo. Usando estos factores, la ANHD espera estimar las causas directas del desplazamiento: precios más altos basados en el comportamiento especulativo y una afluencia de nuevos residentes (ANHD, 2015).

3.1.1. DAP Map

La Figura 3.1 muestra el mapa interactivo (llamado DAP Map - *Displacement Alert Project Map*) que incluye siete capas de visualización.

Tres de estas capas muestran el nivel de agregación geográfica de la información en cuanto a la región de NYC: Distrito del Consejo, Distrito Comunitario y Código Postal. Cada capa cuenta con un menú desplegable para acercar a una área especifica. El usuario puede encontrar más información sobre las métricas de un edificio en particular haciendo click en el lote del edificio.

DAP Map tiene cuatro métricas contenidas en el mapa de visualización de datos. El DAP Map muestra las siguientes capas (ANHD, 2015):

- Pérdida de unidades rentadas en el edificio.
- Volumen de permisos de construcción del Departamento de Edificios de la Ciudad de Nueva York.
- Precio de venta de la construcción.
- Indicador combinado de estabilización de renta y vulnerabilidad del inquilino.

Cada métrica anterior se muestra para cada edificio de NYC, y contiene una medida de riesgo que esta codificado en un gradiente de color. Este gradiente de color indica el nivel de riesgo de desplazamiento de cada edificio. Se muestra una escala de color identificando con el color rojo a aquellos edificios de alto riesgo y con amarillo a los de bajo riesgo

Las cuatro métricas mencionadas previamente han sido calculados de la siguiente manera:

Figura 3.1: DAP Map (ANHD, 2015)

Pérdida de unidades rentadas
Esta capa muestra el porcentaje de pérdidas de unidades rentadas en cada edificio entre 2007 y 2014, con colores más oscuros que indican una mayor pérdida. Esta capa incluye todos los edificios de departamentos con al menos una unidad de renta entre 2007 y 2014 en edificios con seis o más unidades residenciales.
Fuente de datos: taxbills.nyc (github.com/talos, CC BY-SA)

Volumen de solicitudes de permiso de construcción
Las solicitudes de permisos de construcción muestran la intención del propietario de hacer construcciones y renovaciones. Por ejemplo, los permisos 'Alt 2' son a menudo renovaciones en departamentos individuales que ocurren después de que ha sido desocupado y pueden ser utilizados para aumentar significativamente la renta.
Esta capa muestra el número de trabajos de construcción del NYC Departament (con permisos tipos: Alt 1, Alt 2, Alt 3 y Demolición) en cada edificio desde 2013; con colores más oscuros que indican más permisos archivados.
Alt 1 y los permisos de demolición son más pesados, seguidos por permiso Alt 2 .
Esta capa incluye todos los edificios residenciales con solicitudes de permisos desde 2013 hasta 2015 para edificios con seis o más unidades residenciales.
Fuente de datos: DOB Permit Issuance and Historical DOB Permit Issuance from the NYC Open Data Portal

Venta de propiedades
Esta capa indica si el precio de venta del edificio podría indicar una estrategia de inversión especulativa. Se muestra las ventas de edificios en 2015; un alto precio de venta significa que el nuevo propietario planea desplazar a los inquilinos existentes para obtener grandes ganancias.
Fuente de datos: DOF Rolling/Annualized Sales

Riesgo combinado de construcción
Esta capa muestra la medida de riesgo del edificio, que indica si el precio de venta, la pérdida de unidades de renta y el número de solicitudes de permisos de construcción sugieren que las viviendas asequibles en este edificio tienen un riesgo alto.
El mapa incluye todos los edificios de departamentos con seis o más unidades residenciales que tienen al menos uno de los tres indicadores de desplazamiento anteriores. Cada categoría de indicador se puntúa de 0 a 100. Se combinan para que un edificio con una gran pérdida de unidades rentadas, un gran volumen de solicitudes de permisos y un alto precio de venta obtenga un puntaje de 300. Un edificio que solo tenía 1 indicador, es decir, solo permisos DOB, pero no venta de propiedad ni pérdida de unidades rentadas: solo puntuaría hasta 100 puntos.

3.2. Proyecto de desplazamiento urbano

Proyecto de desplazamiento urbano (Urban Displacement Project) es una iniciativa de investigación de UC Berkeley en colaboración con investigadores de la UCLA y Portland State University con aportes y fondos de organizaciones comunitarias, agencias de planificación regional, fundaciones y la Junta de Recursos del Aire del Estado de California (ARB) (UCB & UCLA, 2017). El objetivo del proyecto es comprender la naturaleza de la gentrificación y el desplazamiento en las ciudades estadounidenses, con un enfoque actual en el área de la bahía de San Francisco y el sur de California.

Las regiones de todo California comenzaron a implementar sus Estrategias de Comunidades Sustentables (SCS, por sus siglas en inglés) en cumplimiento con el Proyecto de Ley Senatorial 375 para implementar la ley estatal de calentamiento global (AB32 de 2006). Desde entonces, las comunidades se han preocupado cada vez más acerca de cómo la nueva inversión de tránsito afectarán las vidas de los residentes existentes, particularmente las comunidades de bajos ingresos y las comunidades de color.

Con el financiamiento de la Junta de Recursos del Aire de California, así como de California Endowment, Urban Displacement Project responde a estas preocupaciones examinando las relaciones entre inversión, cambio del vecindario, gentrificación y desplazamiento.

Se ha realizado una investigación detallada del estudio de la interacción de estos factores, y se ha publicado los siguientes artículos (UCB & UCLA, 2017):

- Race, Displacement, and Fair Housing
- Research Brief on the Consequences of Displacement
- Developing a new methodology for analyzing potential displacement
- Research Brief on Housing Production and Displacement
- Literature Review on Gentrification and Displacement
- Displacement Maps
- Previous Work: Susceptibility Report
- Case studies on gentrification, displacement, and development impact

A partir de estos artículos concluyen que la gentrificación, o la afluencia de capital y de la población con mayores ingresos, así como la presencia de residentes de educación superior en barrios de clase trabajadora ya ha transformado alrededor del 10 % de los barrios del Área de la Bahía de SF (UCB & UCLA, 2017). El desplazamiento que ocurre cuando la vivienda o las condiciones del vecindario hacen presión, está ocurriendo en el 48 % de los barrios del Área de la Bahía, divididos casi de manera uniforme entre los barrios de bajos ingresos y de ingresos moderados / altos (UCB & UCLA, 2017). El desplazamiento puede ser físico (deterioro de condiciones de vivienda) o económicas (a medida que aumentan los costos). Se puede expulsar de los hogares, o se puede prohibir que se muden allí. A esto se le llama desplazamiento excluyente (UCB & UCLA, 2017). El desplazamiento, ya sea físico o económico, puede resultar tanto de la desinversión como de la inversión. Por lo tanto, el desplazamiento a menudo se lleva a cabo con la gentrificación, en ninguna parte a plena vista. Varios factores clave están detrás de la gentrificación y el desplazamiento: proximidad a las estaciones de tren, centros de trabajo , barrios históricos, y un fuerte real state market (UCB & UCLA, 2017).

Extendiendo estos hallazgos aún más, se ha profundizado en nueve barrios del Área de la Bahía para rastrear la trayectoria de gentrificación y respuesta de la comunidad. Se ha encontrado que (UCB & UCLA, 2017):

- La gentrificación puede no preceder al desplazamiento. A menudo se supone que la gentrificación es un precursor al desplazamiento residencial. Sin embargo, en muchos de sus casos se encuentra que el desplazamiento precede a la gentrificación y que los dos procesos a menudo ocurren simultáneamente.

- La gentrificación y el desplazamiento son regionales. A pesar de que la gentrificación y el desplazamiento son a menudo visto como un fenómeno local o de barrio, los casos muestran que están inherentemente vinculados a cambios en el mercado regional de vivienda y trabajo.

- A pesar de las continuas presiones y ansiedad, muchos barrios que esperaban estar en riesgo de desplazamiento - como East Palo Alto, Marin Ciudad y Chinatown de San Francisco - han sido sorprendentemente estables, al menos hasta 2013, el más reciente año con datos disponibles. Esto probablemente se deba a una combinación de producción de viviendas subsidiadas, protecciones de inquilinos, control de alquileres y una organización fuerte de la comunidad.

- Las políticas, la planificación y la organización pueden estabilizar barrios. Muchos de los casos han demostrado estabilidad notable, en gran parte debido a las políticas de vivienda local, organización comunitaria, protección del inquilino y técnicas de planificación.

- La inversión en transporte da forma al desplazamiento. La investigación sugiere que no son solo las inversiones en el transporte y la infraestructura que puede acelerar los procesos de gentrificación y desplazamiento, sino también la planificación de tales inversiones.

Urban Project Displacement ofrece una guía interactiva para la gentrificación y el desplazamiento en cada vecindario en el Área de la Bahía y Los Ángeles, en forma de mapa (Figuras 3.2 y 3.3). El mapa funciona como un sistema de alerta temprana regional en el nivel de zona censal, con clasificaciones que varían desde no perder viviendas de bajos ingresos a gentrificación avanzada y exclusión avanzada de viviendas de bajos ingresos.

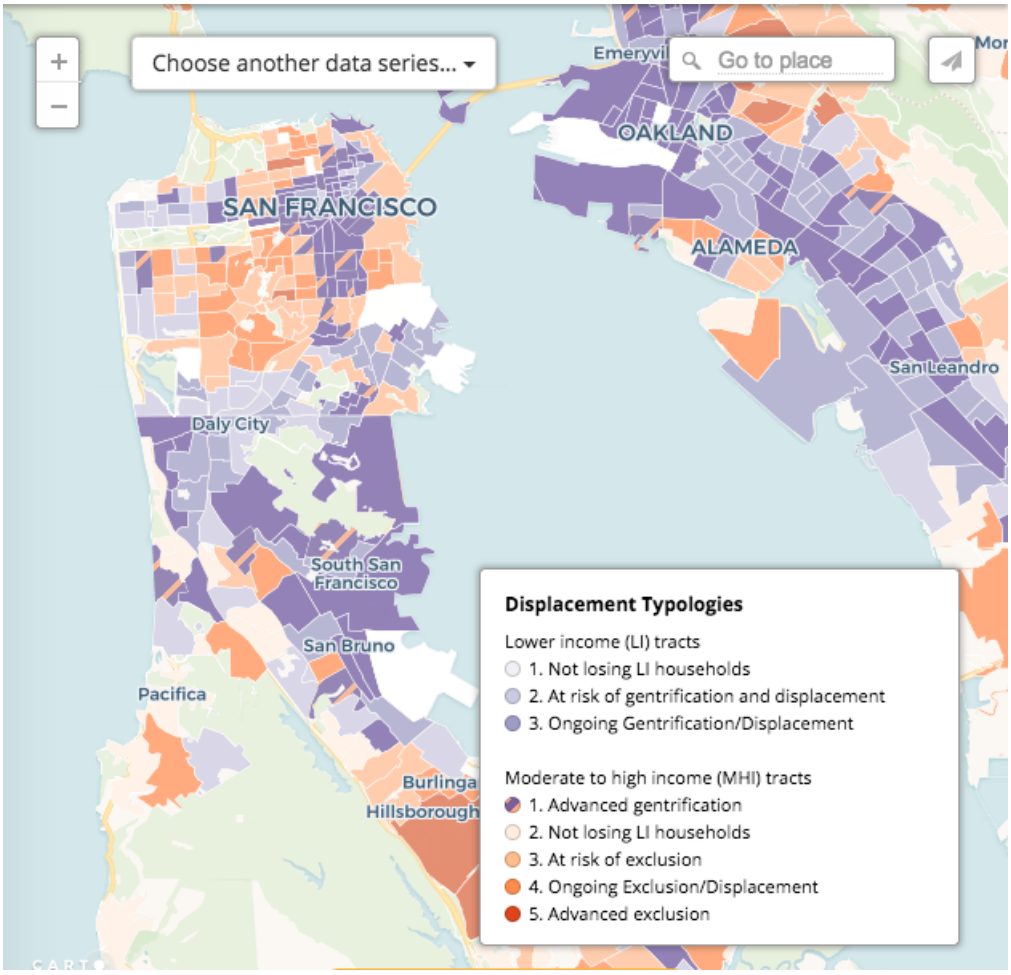

Figura 3.2: *Urban Displacement Project SF* (UCB & UCLA, 2017)

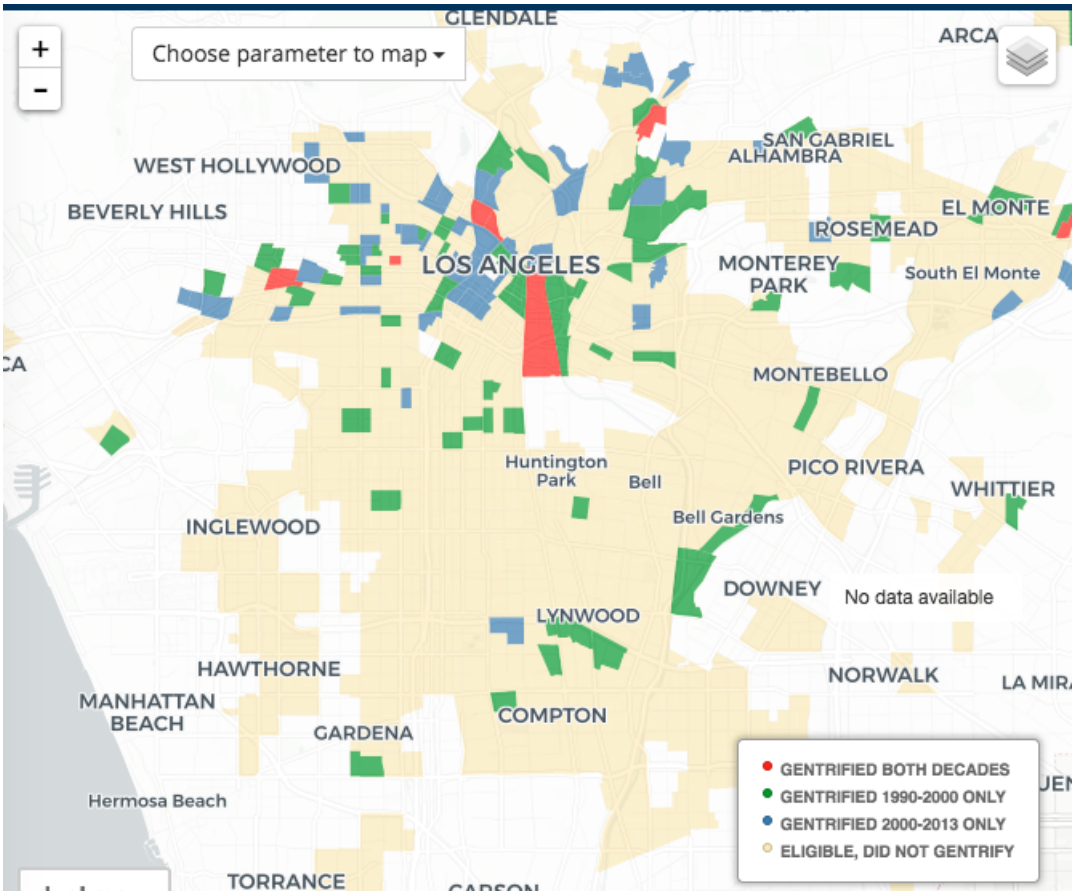

Figura 3.3: *Urban Displacement Project LA* (UCB & UCLA, 2017)

Metodología

Se han analizado más de 50 variables en el periodo 1990-2013 de varios conjuntos de datos que incluyen datos demográficos, transporte, vivienda, uso de la tierra y políticas. Se desarrolla un índice de gentrificación para caracterizar lugares que históricamente albergan poblaciones vulnerables y han experimentado importantes cambios demográficos e inversión en bienes raíces (UCB & UCLA, 2017).

Para aproximar el desplazamiento, se calcula la pérdida de hogares de bajos ingresos para cada período de tiempo. Algunos investigadores han encontrado que la composición de barrios en los Estados Unidos es considerablemente estable; por lo tanto, se supone que cualquier vecindario que experimenta una pérdida de hogares de bajos ingresos es resultado de las presiones de desplazamiento.

Aunque el cambio en los hogares de bajos ingresos podría deberse a la movilidad de los ingresos (es decir, los hogares de bajos ingresos que se mueven a las categorías de ingresos medios y altos, o viceversa), en el análisis de datos del Panel de Estudio sobre Dinámica de Ingresos se estima que habrá un aumento neto en el número de hogares de bajos ingresos en la mayoría de los lugares. Por lo tanto, las estimaciones de desplazamiento probablemente estén subestimadas. Se han construido modelos de regresión para estimar los predictores de la gentrificación y la pérdida del número de hogares de bajos ingresos / desplazamiento, las cuales fueron incorporadas en las categorías de lugares para el riesgo de desplazamiento relacionado con la gentrificación o desplazamiento de exclusión que ocurre en barrios de mayores ingresos. Se concluye que la proximidad a una estación de ferrocarril afecta los patrones de cambio de vecindario (*neighborhood change*) asociado con la gentrificación y el desplazamiento. Como parte del proceso de tratar de modelar el cambio de vecindario, se realiza un modelo de regresión logística que permite comprender mejor la relación entre barrios de tránsito, gentrificación y desplazamiento

3.2.1. Modelado del desplazamiento del vecindario: Gentrificación

Se construye un índice de gentrificación mediante el ajuste de modelos de regresión logística. Como primer paso se define los criterios que debe cumplir un vecindario (zona censal) para estar gentrificado entre los años 1 y 2 (UCB & UCLA, 2017).

Para Los Angeles, una zona es vulnerable a la gentrificación (o elegible para gentrificar) si cumple todos los siguientes criterios:

- La zona censal tiene una población de al menos 500 residentes en el año 1
- Vulnerable, cumpliendo 3 de 4 de los siguientes indicadores:
  a) el porcentaje de hogares de bajos ingresos (cuyo ingreso familiar sea inferior al 80 % del ingreso medio del condado) sea mayor que la mediana del condado
  b) el porcentaje de población con educación universitaria es menor que la mediana poblacional del condado;
  c) el porcentaje de inquilinos es mayor que la mediana del condado y
  d) el porcentaje de población “no blanca” es mayor que el mediana poblacional del condado

Se dice que la zona está gentrificada o gentrificándose si cumple con los requisitos de elegibilidad y todos los siguientes criterios:

- Cambio demográfico entre los años 1 y 2:
  a) Cambio porcentual de población con educación universitaria es mayor que la del condado;
  b) Cambio porcentual de población blanca no hispana es mayor que la del condado;
  d) Cambio porcentual del ingreso medio del hogar es mayor que la del condado

- Cambio en la renta media bruta siendo mayor que la renta bruta media del condado

Utilizando la definición anterior para Los Angeles, se ha encontrado que 81 zonas censales se gentrificaron entre 1990 y 2000,y 82 zonas censales se gentrificaron entre los años 2000 y 2013. De esos 82 zonas que se gentrificaron entre 2000 y 2013, ocho también se habían gentrificado en la década anterior. Se estima que en total 155 zonas censales se gentrificaron entre 1990 y 2013 en Los Angeles (UCB & UCLA, 2017).

Los tramos gentrificados se muestran en la Figura 3.3. Incluye tramos gentrificados en cada período de tiempo y aquellos que gentrificaron en ambos períodos de tiempo. También se muestran los tramos vulnerables.

Para el Área de la Bahía de San Francisco, el índice se modifica ligeramente para reflejar las condiciones de la región. Primeramente las medidas se comparan con la mediana regional que incluye nueve condados. Segundo, no se usa los cambios de los blancos no hispanos en los criterios de cambio demográfico, ya que una considerable investigación ha surgido sobre la naturaleza de la gentrificación impulsada por negros y asiáticos en mercados fuertes como el Área de la Bahía de SF. Finalmente, debido al papel de la afluencia de capital global en el mercado de la vivienda, se utiliza una combinación de incrementos en los precios de la vivienda y unidades nuevas en el market-rate para el segundo criterio de cambio (UCB & UCLA, 2017).

Para el Área de la Bahía, una zona es vulnerable a la gentrificación si cumple con todos los siguientes criterios:

- El tramo (*track*) tenía una población de al menos 500 residentes en el año 1

- Vulnerable, cumpliendo 3 de 4 de los siguientes indicadores:
  a) el porcentaje de hogares de bajos ingresos (cuyo ingreso familiar sea inferior al 80 % del ingreso medio del condado) sea mayor que la mediana del condado
  b) el porcentaje de población con educación universitaria es menor que la

mediana poblacional del condado;
c) el porcentaje de inquilinos es mayor que la mediana del condado y
d) el porcentaje de población "no blanca" es mayor que el mediana poblacional del condado

Se dice que una zona está gentrificada o gentrificadose si cumple con los requisitos de elegibilidad y todos los siguientes criterios:

- Cambio demográfico entre los años 1 y 2:
  a) Cambio porcentual de población con educación universitaria es mayor que la región
  b) Cambio porcentual del ingreso medio del hogar es mayor que la del condado

- Inversión entre los años 1 y 2:
  a) el porcentaje de unidades construidas y la tasa de mercado es mayor a la mediana regional
  b) Crecimiento en los siguientes criterios :
  i)el porcentaje de aumento del precio de venta de una vivienda "*single-family*" por pie cuadrado es mayor que la mediana regional ;
  ii) el porcentaje de aumento del precio de venta de viviendas "*multi-family*" por pie cuadrado es mayor a la mediana regional y
  iii)porcentaje de aumento del valor de la casa es mayor que la mediana regional.

Utilizando los criterios anteriores para el Área de la Bahía de SF, se ha encontrado que 83 zonas censales han sido gentrificadas entre 1990 y 2000 y 85 zonas han sido gentrificadas entre los años 2000 y 2013. En total se estima que 149 zonas han sido gentrificadas entre 1990 y 2013. (UCB & UCLA, 2017)

El hecho de que una zona se haya gentrificado entre dos años no les impide continuar el cambio. De hecho, de los 149 zonas que se estiman haber gentrificado entre 1990 y 2013, 71 tienen tasas más bajas de crecimiento de hogares de bajos ingresos que el resto de la región, 105 han perdido viviendas asequibles, y 100 han tenido tasas menores de inmigración de residentes de bajos ingresos en 2013 que en 2009. Además, 88 de los tramos gentrificados continúan teniendo una mayor proporción de hogares de bajos ingresos que la región (39 %) (UCB & UCLA, 2017).

Regresión Logística

La gentrificación puede incluir tanto el desplazamiento directo (desventaja social y económica de residentes que son expulsados) y el desplazamiento excluyente

(barreras que dificultan que los residentes desfavorecidos se muden). Es difícil separar estos dos elementos del modelo de regresión. Se modela la gentrificación por dos períodos de tiempo individuales: 1990-2000 y 2000-2013 (UCB & UCLA, 2017).

Para Los Ángeles y el Área de la Bahía de SF, se utiliza un modelo de regresión logística con dos tipos de resultados de regresión (Figuras 3.4 y 3.5). Los primeros dos modelos (I y II) solo miran las zonas que son elegibles para gentrificación, mientras que el segundo conjunto de modelos analiza todas las zonas (III y IV). La variable dependiente es una variable dicotómica que indica si un tramo se ha gentrificado o no. Las variables independientes incluyen factores clave relacionados con la gentrificación (raza e ingresos), una variable de tenencia (porcentaje de inquilinos), y dos variables de lugar (barrios de tránsito y vecindarios de tránsito del centro, etiquetados TOD).

En este análisis, se separan barrios de tránsito en tres categorías dependiendo del año en que se ha abierto la estación de tránsito: barrios de tránsito de la década de 1990 (abiertos en la década de 1990), vecindarios de tránsito de la década de 2000 (la estación de tren se abrió en la década de 2000) y barrios de tránsito recientes (la estación de tren abrió en 2012 o más tarde solo para Los Ángeles, ya que hay una gran cantidad de desarrollos de estaciones recientes en LA comparado con el Área de la Bahía de SF).

Además, se incluye una variable de entorno incorporado (Porcentaje de unidades de vivienda construidos antes de la Segunda Guerra Mundial, definidos como aquellos que fueron construidos antes de 1950) y una variable de accesibilidad (número de trabajos por milla cuadrada). Los datos del año base para las variables independientes son ya sea 1990 o 2000 dependiendo del período examinado.

	Eligible Tracts				*All Tracts*			
	Model I BA		**Model II BA**		**Model III BA**		**Model IV BA**	
	1990-2000		2000-2013		1990-2000		2000-2013	
Intercept	-6.690	***	-4.861	***	-8.060	***	-7.191	***
Median Household Income (/10000)	0.692	**	0.332		0.765	**	0.698	**
Income Squared	-0.032		-0.011		-0.059	**	-0.057	**
% non-Hispanic black	0.012		2.030	**	1.383	*	3.772	***
% Asian	-0.890		-0.362		0.256		1.385	
% Hispanic	-0.711		-0.242		1.800	**	2.216	***
% Renters	2.373	***	0.598		3.524	***	1.412	*
Downtown TOD	1.906	***	0.782	**	1.363	***	0.366	
Non-Downtown TOD	0.841	**	-0.269		1.058	***	0.087	
TOD 1990s	0.823	**	-0.465		0.883	***	-0.179	
TOD 2000s	-		0.354		-		0.372	
% of Housing Units Prewar (<1950)	0.438		1.783	***	-0.143		1.039	*
Employment Density (# jobs / square mile)	0.000		0.000		0.000		0.000	*
N	640		626		1576		1579	
Likelihood Ratio	219.9	***	229.9	***	262.5	***	266.7	***
***<.01 **<.05 *<.10								

Source: 1990 and 2000 Decennial Censuses, 2009-13 5-year ACS
Tabulations by M. Zuk Aug 2015

Figura 3.4: Modelos de regresión logística para el Área de la Bahía de SF (UCB & UCLA, 2017)

	Eligible Tracts				*All Tracts*			
	Model I LA		**Model II LA**		**Model III LA**		**Model IV LA**	
	1990-2000		2000-2013		1990-2000		2000-2013	
Intercept	-3.2807	***	2.6899	***	-5.7477	***	-4.5411	***
Median Household Income (/10000)	-0.2130	**	-0.8161	***	0.4623	***	0.2741	***
Income Squared	0.0208	*	0.0852	***	-0.0111	***	-0.0240	***
% non-Hispanic black	0.0065	***	-0.0756	***	-0.0069	***	-0.0124	***
% Asian	0.0273	***	-0.0296	***	-0.0157	***	0.0015	
% Hispanic	0.0126	***	-0.0538	***	-0.0106	***	-0.0160	***
% Renters	-0.0065	***	0.0026		0.0214	***	0.0247	***
Downtown TOD	0.5736	***	0.4838	***	0.7406	***	0.6822	***
TOD 1990s	0.1327	**	-0.0381		0.3575	***	-0.0193	
TOD 2000s	-		-0.2962	***	-		-0.2677	***
TOD Recent	-		1.0297	***	-		0.3971	***
% of Housing Units Prewar (<1950)	0.0178	***	0.0345	***	0.0259	***	0.0309	***
Employment Density (# jobs / square mile)	0.0001	***	0.0006	***	0.0001	***	0.0002	***
N	937		929		2,273		2,306	
Likelihood Ratio	493.110	***	2157.547	***	7822.79	***	6436.391	***
***<.01 **<.05 *<.10								

Source: 1990 and 2000 Decennial Censuses, 2009-13 5-year ACS, NETS (1990, 2000)
Tabulations by C.Pech & P. Ong, July 2015

Figura 3.5: Modelos de regresión logística para LA (UCB & UCLA, 2017)

3.3. Resumen

El Dap Map es una herramienta de visualización de NYC que permite observar las siguientes características (ANHD, 2015):

- Pérdida de unidades rentadas en el edificio.
- Volumen de permisos de construcción del Departamento de Edificios de la Ciudad de Nueva York.
- Precio de venta de la construcción.
- Indicador combinado de estabilización de renta y vulnerabilidad del inquilino.

Estas características son consideradas con un rol importante en el desplazamiento en NYC, y finalmente se calcula una métrica que evalúa el riesgo de desplazamiento como una ponderación de las características antes mencionadas.

Por otro lado, el Proyecto de Desplazamiento Urbano realiza una análisis para LA y SF mediante el ajuste de un modelo de regresión logística que permite estimar la variable objetivo "gentrificación" o "no gentrificación" , utilizando 12 características (UCB & UCLA, 2017):

- Ingreso medio por vivienda.
- Porcentaje de población blanca no hispana.
- Porcentaje de población asiática.
- Porcentaje de población hispana.
- Porcentaje de inquilinos.
- Vecindarios de tránsito del centro.
- Vecindarios de tránsito que no son del centro.
- Vecindarios de tránsito del centro en 1990.
- Vecindarios de tránsito del centro en el 2000.
- Porcentaje de viviendas construidas antes de 1950.
- Densidad de empleo.

En este caso, estimar la gentrificación no es una ponderación sino más bien tiene un fundamento robusto basado en una regresión logística.

Capítulo 4

El modelado de la gentrificación en CDMX

La gentrificación es un fenómeno que está ocurriendo en las grandes ciudades del mundo. Primeramente hay que estudiar el comportamiento de las zonas de la Ciudad de México (CDMX) para entender cómo está sucediendo la gentrificación en la CDMX. Para realizar dicho análisis se define el método mostrada en la Figura 4.1, el cual está basado en el método CRISP-DM (ver sección 2.2.2). Fue modificado con el objetivo de adaptarlo al problema abordado.

1. Contexto. Esta sección involucra la determinación de objetivos, evaluación de la situación actual ,definición de conceptos utilizados durante el análisis e identificación de fuentes de datos disponibles.

2. Análisis exploratorio de los datos. Esta etapa incluye la recopilación de datos, descripción de datos y verificación de la calidad de los datos. Tiene una estrecha relación con el Contexto debido a que es necesario revisar fuentes de datos de calidad y en caso de que los datos sean insuficientes regresar a la etapa anterior para definir fuentes de datos complementarias.

3. Preparación de los datos. Una vez que las fuentes de datos disponibles son identificadas, es necesario seleccionarla, limpiarla y transformar los datos con el objetivo de prepararlos para el modelado.

4. Modelado. Visualización de datos y ajuste de varias técnicas de modelado. División de los datos en un conjunto de entrenamiento y de prueba. En esta etapa es posible regresar a la etapa de preparación de datos con el objetivo de realizar transformaciones a los datos que sean más fáciles de interpretar por el modelo.

5. Evaluación. Los resultados del modelo deben ser evaluados en el contexto establecido en la primera etapa mediante métricas que miden la precisión de ajuste del modelo para los datos de entrenamiento y prueba.

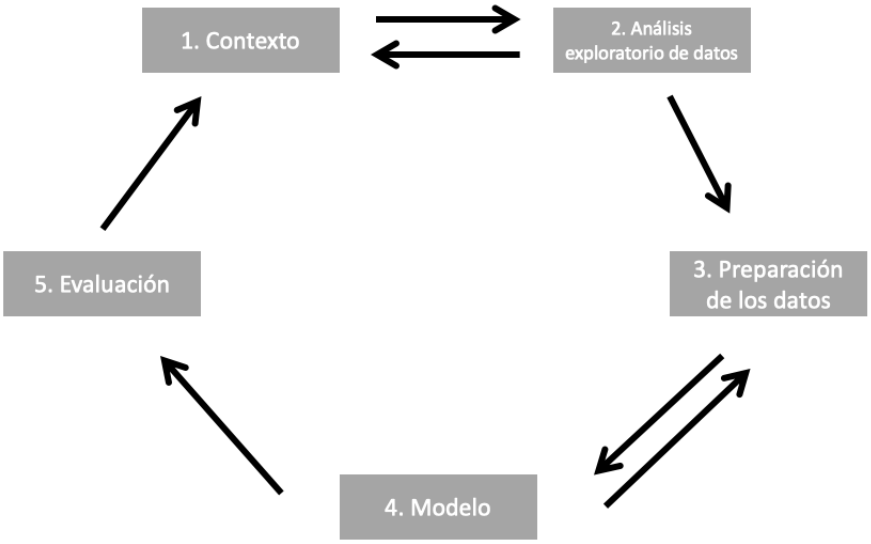

Figura 4.1: Método CRISP-DM

4.1. Contexto

La cobertura geográfica de la presente tesis se limita a la Ciudad de México, la ciudad más poblada de México, ya que es una de las ciudades que ha reportado zonas gentrificadas (Ver Sección 2.6), y existen datos relacionados que permiten modelarla.

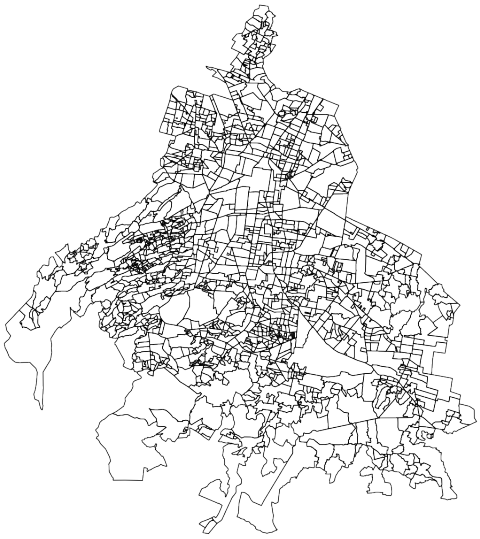

Figura 4.2: Colonias de la CDMX

La figura 4.2 muestra las 1,436 colonias de la CDMX que son manejadas en la presente tesis. La cobertura temporal son los años 2000, 2010 y 2016. Se usan datos para estos años los cuales provienen, en su mayoría, de los Censos de Población y Vivienda realizados por el INEGI.

La Tabla 4.1 muestra las colonias que se definen como gentrificadas en la presente tesis (según (CAMHAJI, 2017), (Forbes, 2014), (Cantera, 2017)) (Ver Sección 2.6).

1.Álamos	9. Irrigación
2.Centro	10. Juárez
3.Condesa	11.Obrera
4.Cuauhtémoc	12. Roma Norte
5.Doctores	13. Roma Sur
6.Escandón	14. San Rafael
7. Hipódromo	15. Santa María la Ribera
8. Hipódromo de la Condesa	16. Tabacalera

Tabla 4.1: Colonias ya gentrificadas

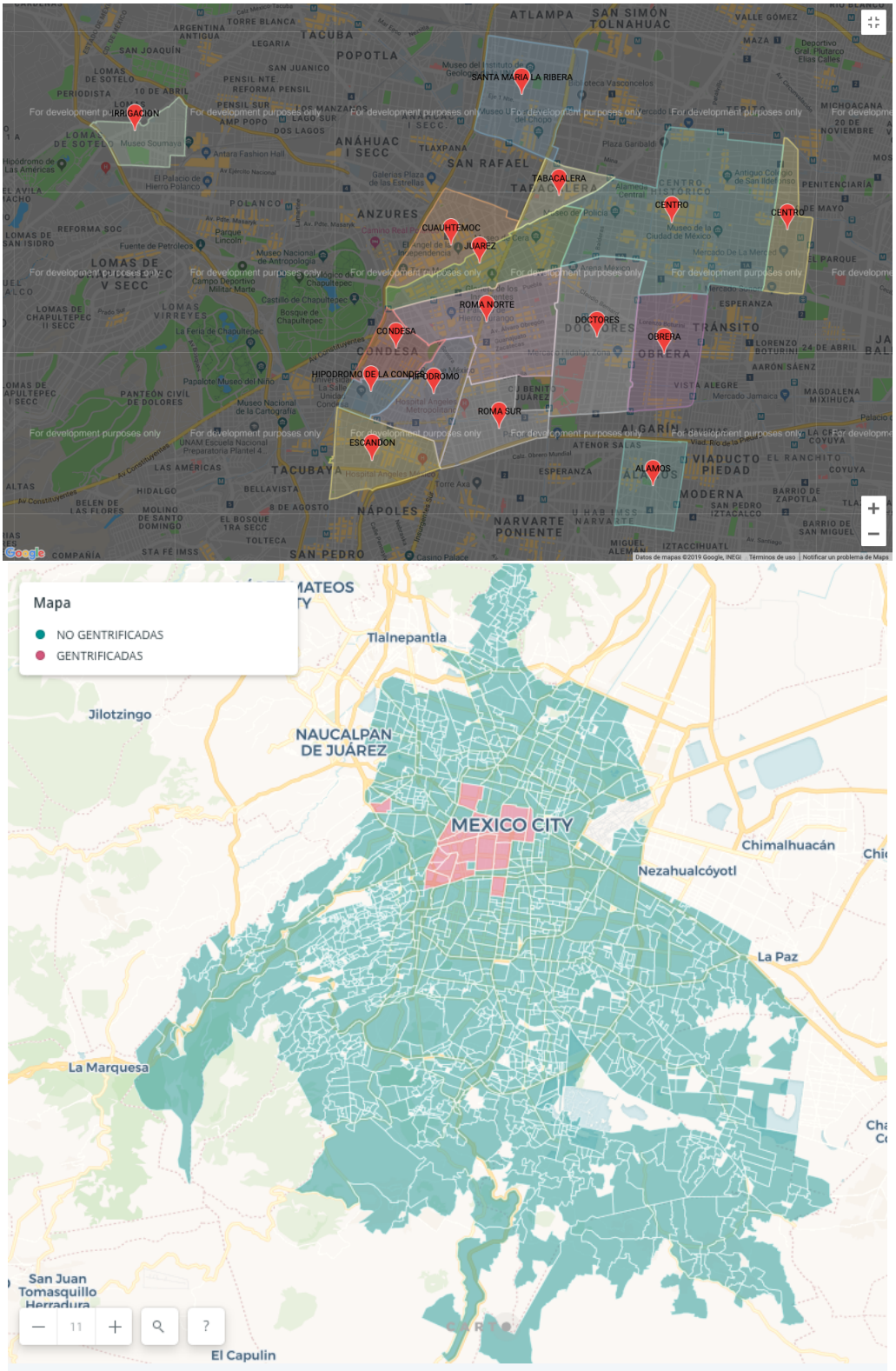

Figura 4.3: Colonias gentrificadas

La Figura 4.3 muestra la ubicación de las colonias gentrificadas sobre un mapa de Google. Se observa que las colonias gentrificadas se encuentran en el centro de la Ciudad de México. En las delegaciones Cuauhtémoc, Miguel Hidalgo y Benito Juárez (Ver Sección 2.6).

El objetivo de esta investigación es proponer un modelo que tenga la capacidad de identificar procesos de gentrificación en la Ciudad de México.

4.2. Análisis exploratorio

En esta sección se realiza un análisis exploratorio de las variables consideradas en la presente tesis a lo largo de toda la Ciudad de México con especial énfasis en la comparación con las colonias gentrificadas. Se define los datos que son ocupados en el presente estudio. Se cuenta con datos provenientes de diversas fuentes: a) censo de población y vivienda del 2000 (SCINCE 2000); b)censo de población y vivienda del 2010, c)Inventario Nacional de viviendas del 2016; d)Directorio Estadístico de Unidades Económicas (DENUE); e) Muestra de datos de precios de venta de departamentos proporcionados por Softec y f)Shapefiles de las delegaciones, Áreas Geoestadísticas Básicas (AGEB), manzanas, localidades y colonias. Las fuentes son mostradas en la Figura 4.4

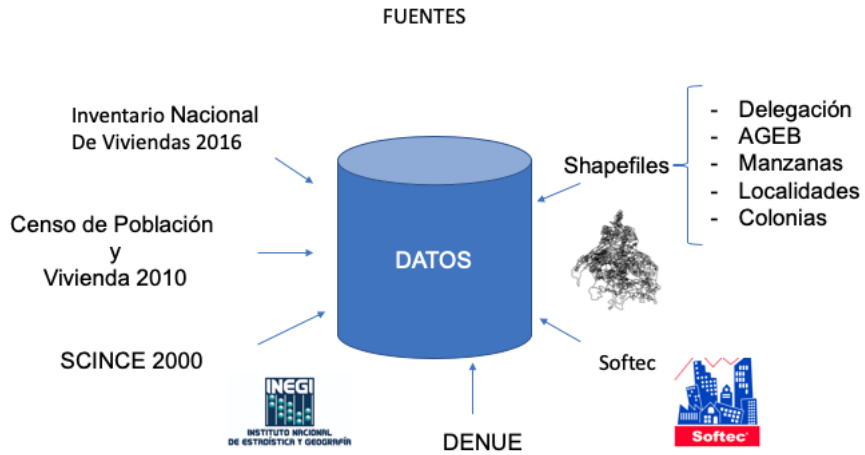

Figura 4.4: Fuentes de datos

Los datos de cada fuente se encuentran con diferentes niveles geográficos de detalle, por lo que se realiza un cruce geográfico de información con la finalidad de estandarizar el nivel de agregación deseado, es decir, a nivel colonia.

Después de realizar el cruce de información, se selecciona las variables que son útiles para el estudio de gentrificación. En la Tabla 4.2 se observa el listado de variables que son analizadas y clasificadas de acuerdo a su categoría.

Categoría	Variables
Socio-demográficas	Porcentaje de población que gana menos de 1 Salario Mínimo(SM) en 2000
	Porcentaje de viviendas rentadas en el 2000
	Porcentaje de viviendas propias y pagadas en el 2000
	Porcentaje de viviendas propias en el 2000
	Porcentaje de población que gana mas de 5 SM en el 2000
	Porcentaje de población que gana de 2 a 5 SM en el 2000
	Porcentaje de población que gana de 1 a 2 SM en el 2000
	Porcentaje de población de 3 años y mas que hablan alguna lengua indígena
	Porcentaje de población que residía en otra entidad en el 2005
	Porcentaje de cambio del promedio de ocupantes por cuarto del 2000 al 2010
	Porcentaje de cambio del total de viviendas habitadas del 2000 al 2010
	Porcentaje de cambio del total de hogares del 2000 al 2010
	Porcentaje de cambio de población soltera del 2000 al 2010
	Porcentaje de cambio de población ocupada del 2000 al 2010
	Porcentaje de cambio del grado promedio de escolaridad del 2000 al 2010
	Porcentaje de cambio de población nacida en otra entidad (2000-2010)
	Porcentaje de cambio de población nacida en la entidad del 2000 al 2010
	Porcentaje de cambio de población de 60 años y mas del 2000 al 2010
	Porcentaje de cambio de población de 15 a 64 años del 2000 al 2010
	Porcentaje de cambio de población de 0 a 14 años del 2000 al 2010
Económicas	Cantidad de lavanderías creadas del periodo 2000 al 2010
	Cantidad de estéticas creadas del periodo 2000 al 2010
	Cantidad de discotecas creadas del periodo 2000 al 2010
	Cantidad de cafeterías creadas del periodo 2000 al 2010
	Cantidad de bares creados del periodo 2000 al 2010
	Cantidad de negocios artísticos creados del periodo 2000 al 2010
	Promedio del valor del metro cuadrado en el 2000
	Promedio del valor del metro cuadrado en el 2010
	Promedio del valor del metro cuadrado en el 2016
	Porcentaje de cambio del valor del metro cuadrado del 2000 al 2010
	Porcentaje de cambio del valor del metro cuadrado del 2010 al 2016
	Porcentaje de cambio del valor del metro cuadrado del 2000 al 2016
Entorno Urbano	Porcentaje de vialidades que no tienen restricción para acceso de personas
	Porcentaje de vialidades que no tienen restricción para acceso de autos
	Porcentaje de vialidades que tienen recubrimiento
	Porcentaje de vialidades que cuenta con señalización
	Porcentaje de vialidades que cuentan con alumbrado público
	Porcentaje de vialidades que cuentan con teléfono público
	Porcentaje de vialidades que cuentan con banquetas
	Porcentaje de vialidades que cuentan con guarniciones
	Porcentaje de vialidades que cuenta con arboles
	Porcentaje de vialidades que cuentan con rampas
	Porcentaje de vialidades que cuentan con puestos semi-fijos
	Porcentaje de vialidades que cuentan con puestos ambulantes

La mayoría de las variables citadas anteriormente provienen de censos, con excepción de las variables relacionadas con el precio del metro cuadrado, que se obtuvo de una muestra de datos de la empresa Softec y se realizo una estimación de dicho valor para el resto de la ciudad. (Ver Sección 4.2.2).

En la Figura 4.5 se muestra la matriz de correlación entre las variables que son usadas como regresores. Se observa que las variables TOTHOG_pctChange (Porcentaje de cambio del total de hogares), TVIVHAB_pctChange (Porcentaje de cambio del totatl de viviendas habitadas), POB0_14_pctChange (Porcentaje de cambio de población de 0 a 14 años) , POB15_64_pctChange (Porcentaje de cambio de población de 15 a 64 años),PNACENT_pctChange (Porcentaje de cambio de ponlación nacida en la CDMX), PNACOE_pctChange (Porcentaje de cambio de población nacida en otra entidad que no es CDMX), POCUPADA_pctChange (Porcentaje de cambio de población ocupada), P12YM_SOLT_pctChange (Porcentaje de cambio de población soltera), GUARNICI_CTV_pct_total (Porcentaje de vialidades con guarniciones) presentan una correlación positiva mayor al 90 % con la variable POBTOT_pctChange (Porcentaje de cambio de población total). Por lo tanto estas variables son eliminadas del listado de variable regresoras.

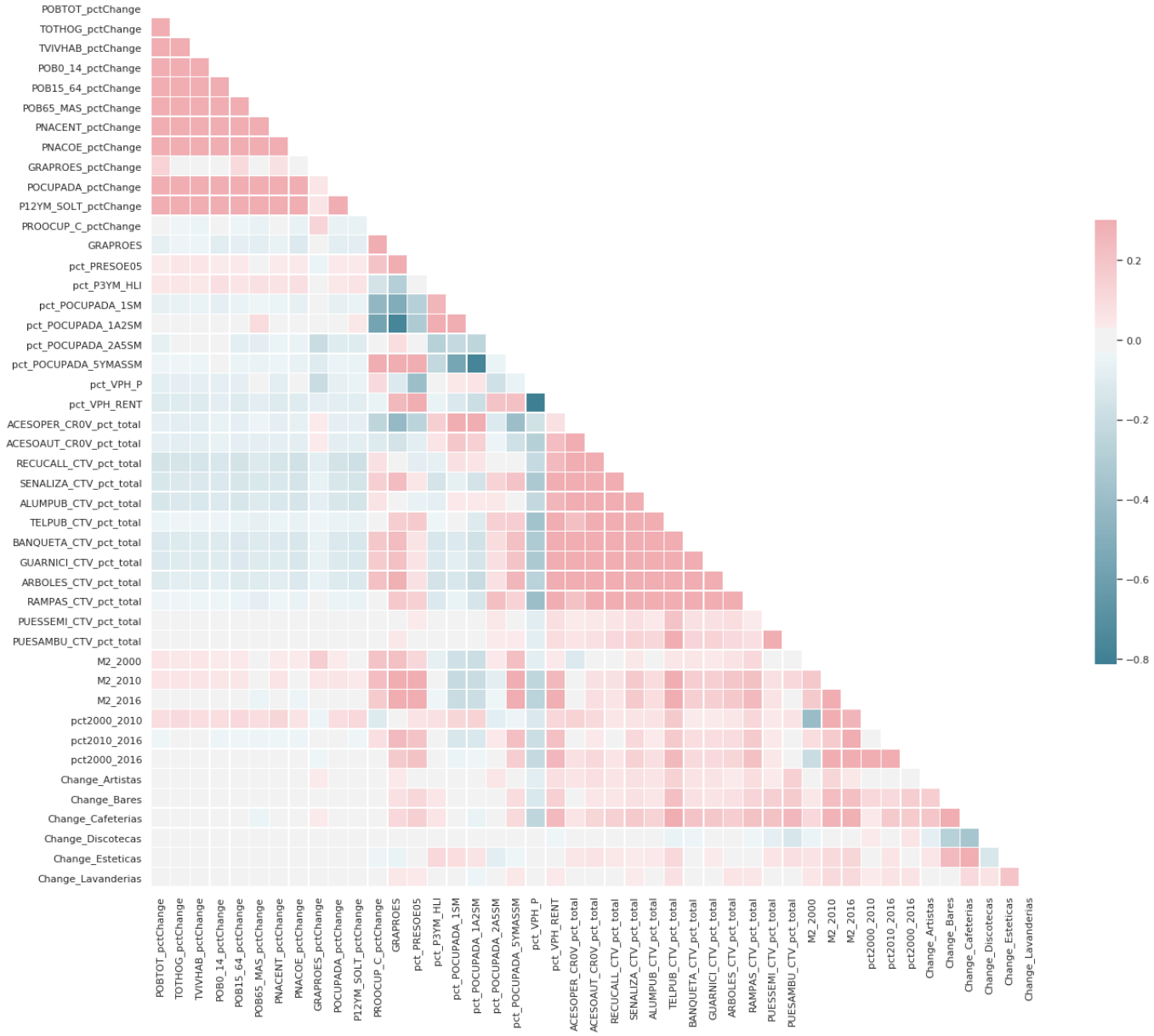

Figura 4.5: Matriz de correlación de las variables

En la Figura 4.6 y 4.7 se muestra el histograma de cada una de las variables. Por ejemplo, para la variable porcentaje de cambio en la población existen colonias que se ubican en la periferia de la ciudad las cuales crecieron significativamente en la cantidad de población en el periodo 2000-2010. En las siguientes secciones se mencionan a estas colonias.

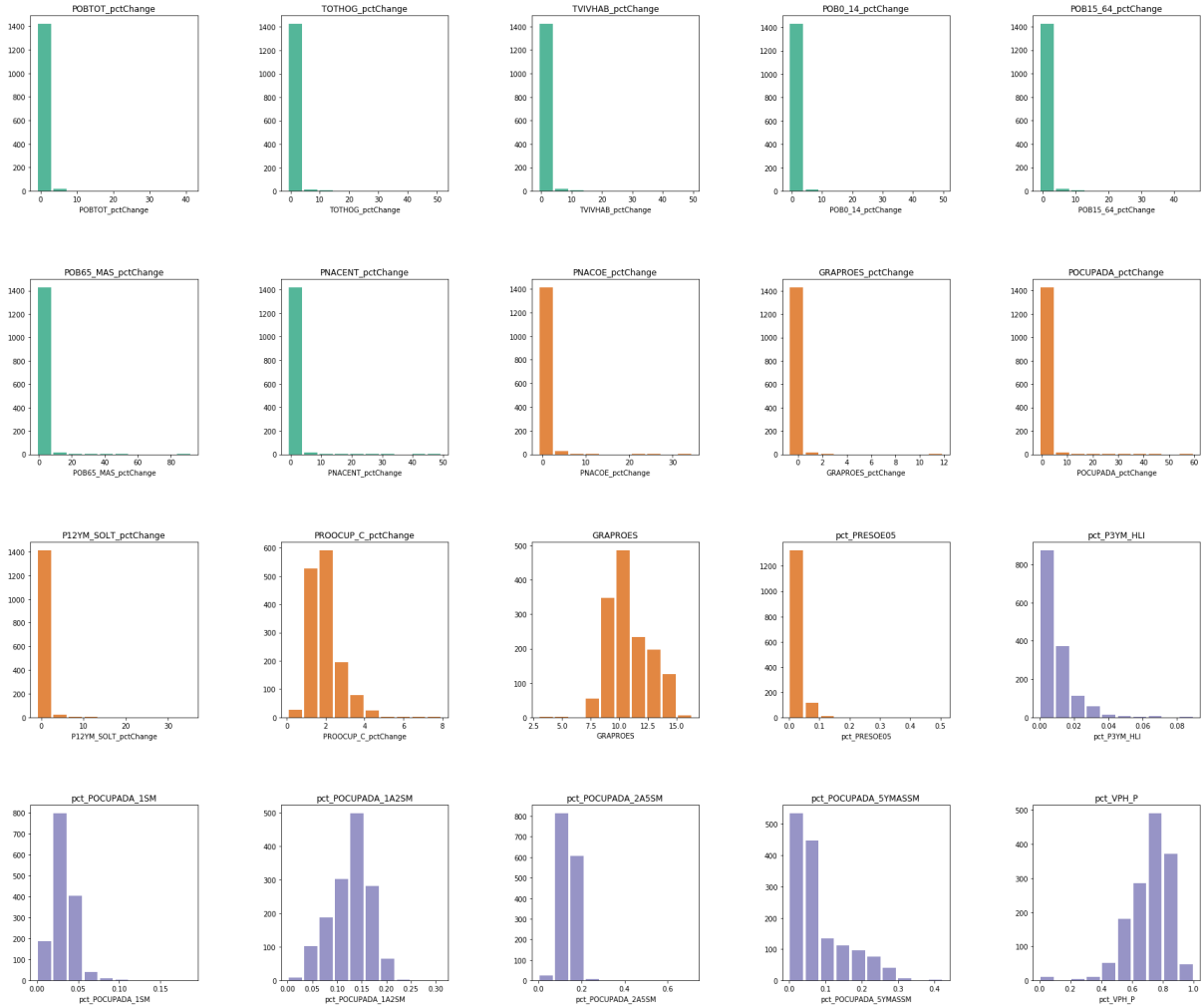

Figura 4.6: Distribución de las variables

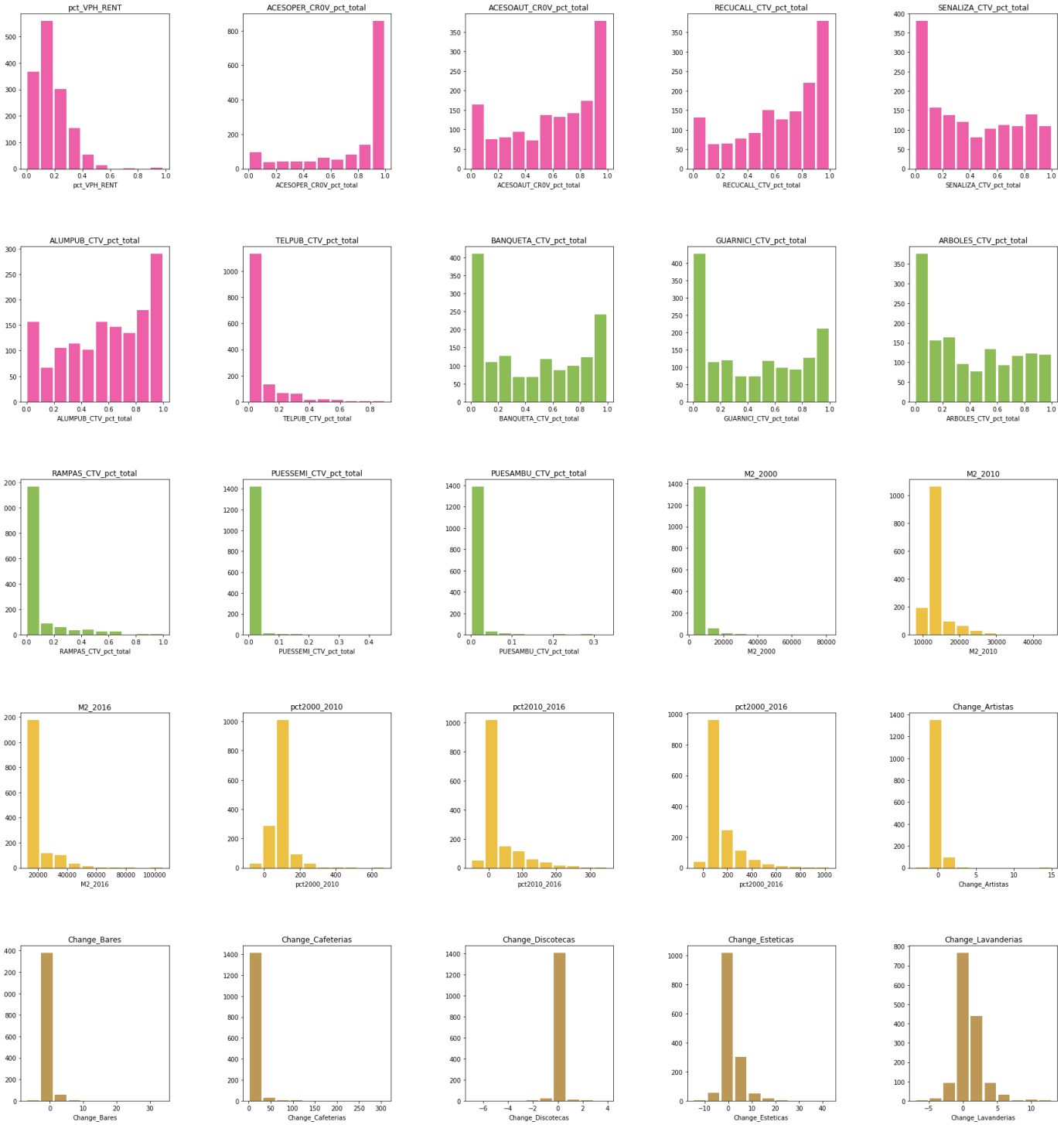

Figura 4.7: Distribución de las variables

Se ha estudiado el comportamiento de las variables para cada colonia y se ha encontrado que las colonias que ya han sido gentrificadas presentan comportamientos diferentes al resto de la ciudad. Algunos ejemplos pueden mostrarse en el siguiente análisis. Las colonias gentrificadas son marcadas con bordes en rojo en los mapas.

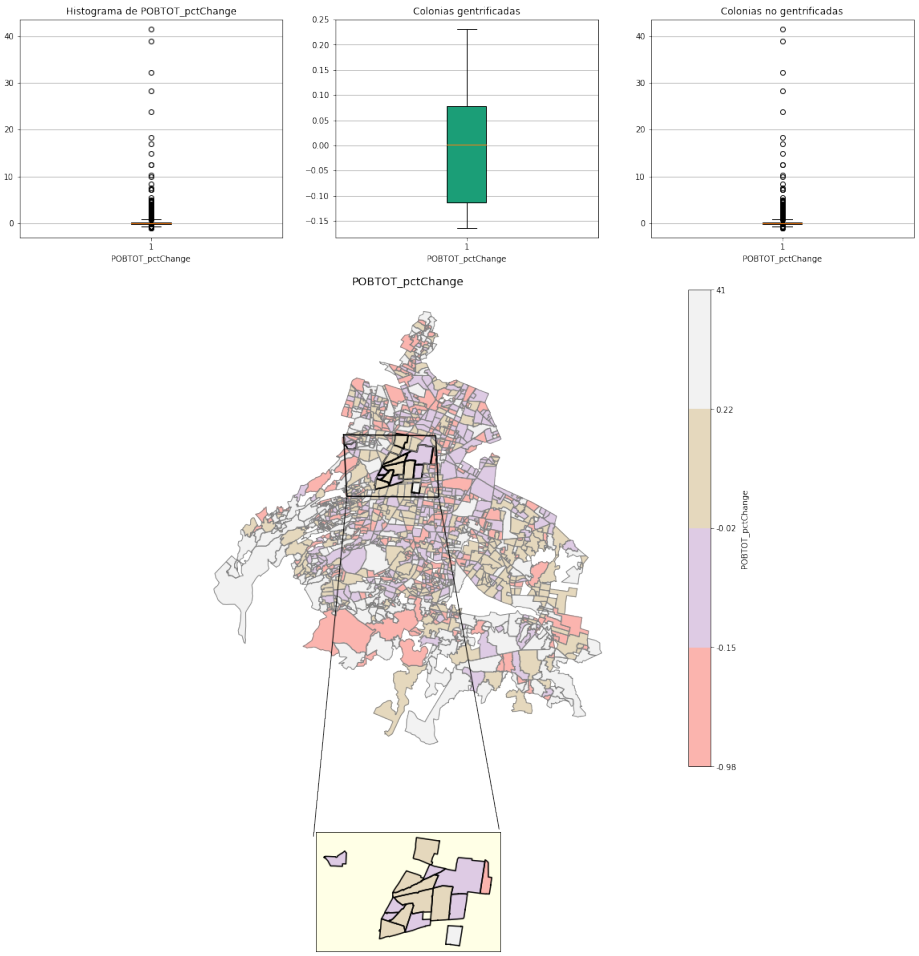

Figura 4.8: Porcentaje de cambio de la población del 2000 al 2010

La Figura 4.8 muestra el cambio porcentual de la población para cada colonia del 2000 al 2016. En el mapa y el diagrama boxplot se muestra para las colonias gentrificadas y los datos restantes. Se observa que el porcentaje de cambio del 25 % de las colonias es menor que -15 % y el 75 % de las colonias tienen un cambio porcentual del 22.48 %. La colonia que tuvo el mayor cambio porcentual negativo es la colonia San Agustín, con un cambio del -97 %, por otro lado se observa la colonia Estación Pantaco la cual muestra un cambio porcentual del 4141 % . En el

mapa se observa con color negro que la periferia de la ciudad tiene valor mayores al 22 %. Es decir, la población en las áreas periféricas de la ciudad ha tenido un aumento significativo en comparación con el centro de la ciudad.

4.2.1. Variables socio-demográficas

En esta sección se muestran los mapas con las variables socio-demográficas que presentan comportamientos relevantes.

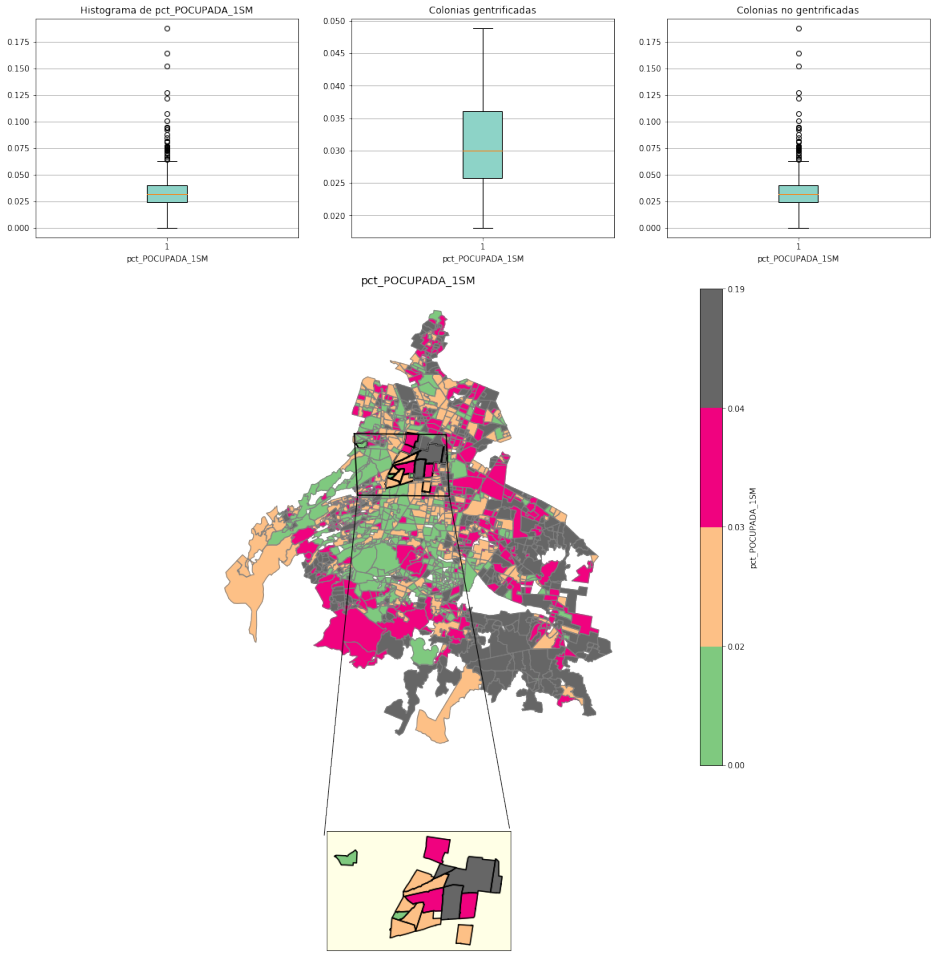

Figura 4.9: Porcentaje de la población que gana menos de 1 salario mínimo en el 2000

La Figura 4.9 muestra el porcentaje de la población a nivel colonia que ganaba menos de 1 salario mínimo (SM) en el año 2000, es decir con nivel socioeconómico bajo. En el mapa se observa que la periferia de la ciudad, en particular, la parte

sur-oriente de la ciudad es la que concentra mayor porcentaje de población con nivel socioeconómico bajo, mayores al 4 % y hasta un 18.75 %. La colonia Bosque de San Juan de Aragón es la que presenta mayor porcentaje de población con esta característica. Por otro lado, las colonias gentrificadas presentan a lo mas 5 % de población que gana menos de 1 SM.

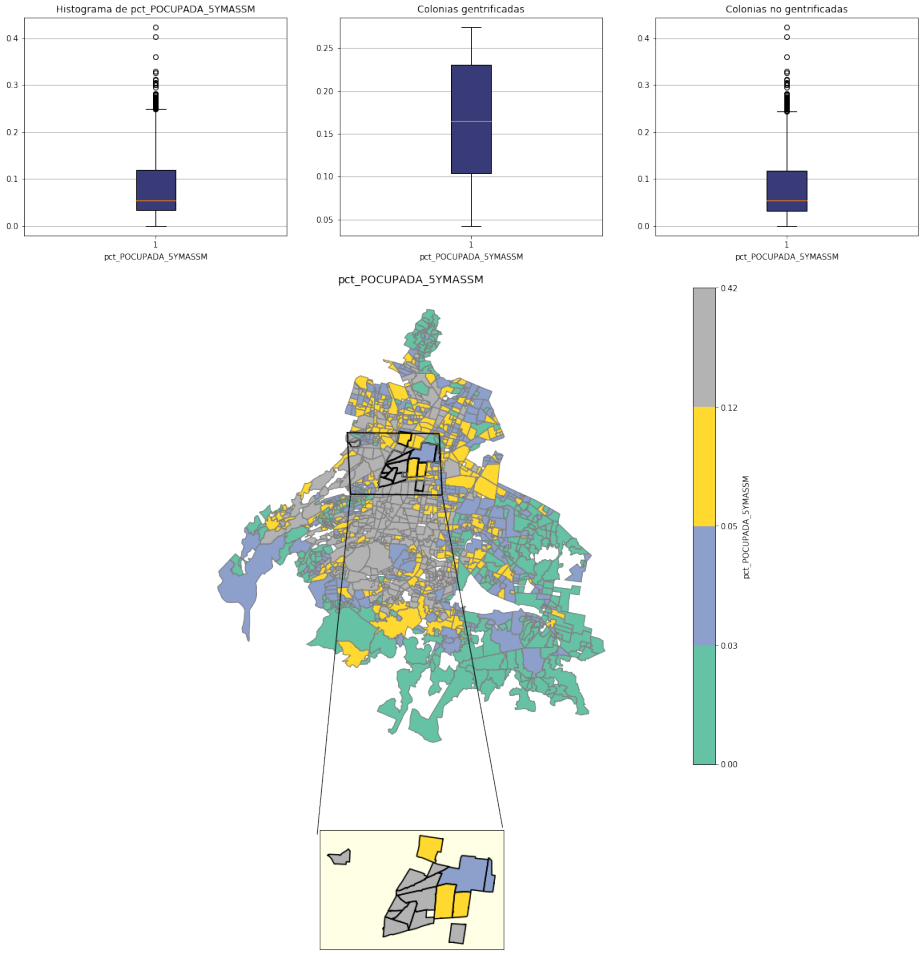

Figura 4.10: Porcentaje de la población que gana mas de 5 salarios mínimo en el 2000

La Figura 4.10 muestra el porcentaje de población a nivel colonia que gana más de 5 salarios mínimos en el año 2000. Se observa que el lado sur-poniente de la ciudad es la zona con población con nivel socio-económico alto que gana mas de 5 salarios mínimos. De las colonias gentrificadas se tiene que el top tres de colonia en esta métrica tienen un 27.42 % , 25.35 % y 25.18 % para las colonias Cuauhtémoc, Hipódromo e Hipódromo de la Condesa, respectivamente. La colonia que tiene el

porcentaje más alto es la colonia Boscoso, con un 42 % de población que gana más de 5 SM.

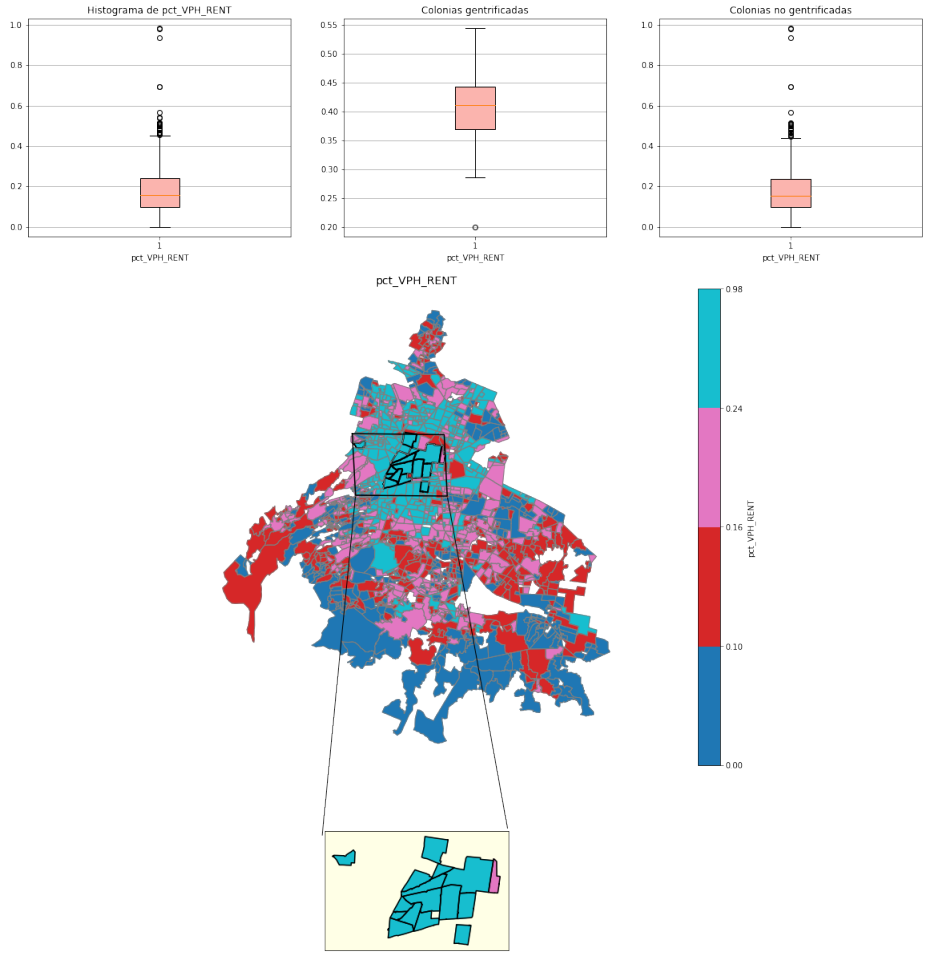

Figura 4.11: Porcentaje de viviendas rentadas en el 2000

La Figura 4.11 muestra el porcentaje de viviendas rentadas a nivel colonia en el año 2000. Se observa que el centro-norte de la ciudad es la zona con mayor porcentaje de viviendas en renta. Las colonias gentrificadas presentan un porcentaje del 20 % al 55 % de viviendas en renta. Por el contrario el sur de la ciudad muestra porcentajes bajos de viviendas en renta, es decir, la mayoría de viviendas en el sur de la ciudad son viviendas propias.

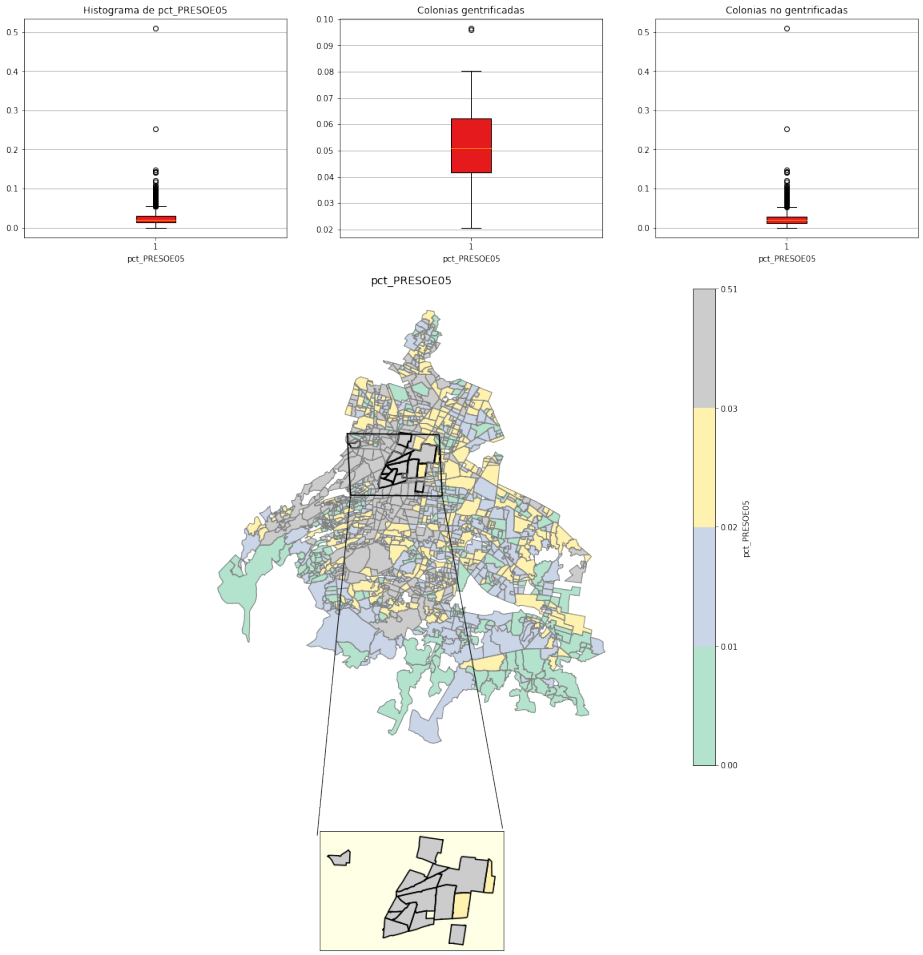

Figura 4.12: Porcentaje de población que residía en otra entidad en el 2005

La Figura 4.12 muestra el porcentaje de la población que residía en otra entidad federativa en el 2005. Esta variable proviene del censo del 2010 y se muestra con la finalidad de tener una métrica relacionada con la migración. Se observa que el poniente de la ciudad es la zona que muestra una mayor rotación de población migratoria al menos a nivel nacional del 2005 al 2010. La colonia gentrificada que presenta un indice bajo de rotación es la colonia Centro con un 2 % de población que emigra a esta zona del 2005 al 2010. Por el contrario, la colonia gentrificada que tiene el mayor índice es la colonia Juárez con un 9.6 %.

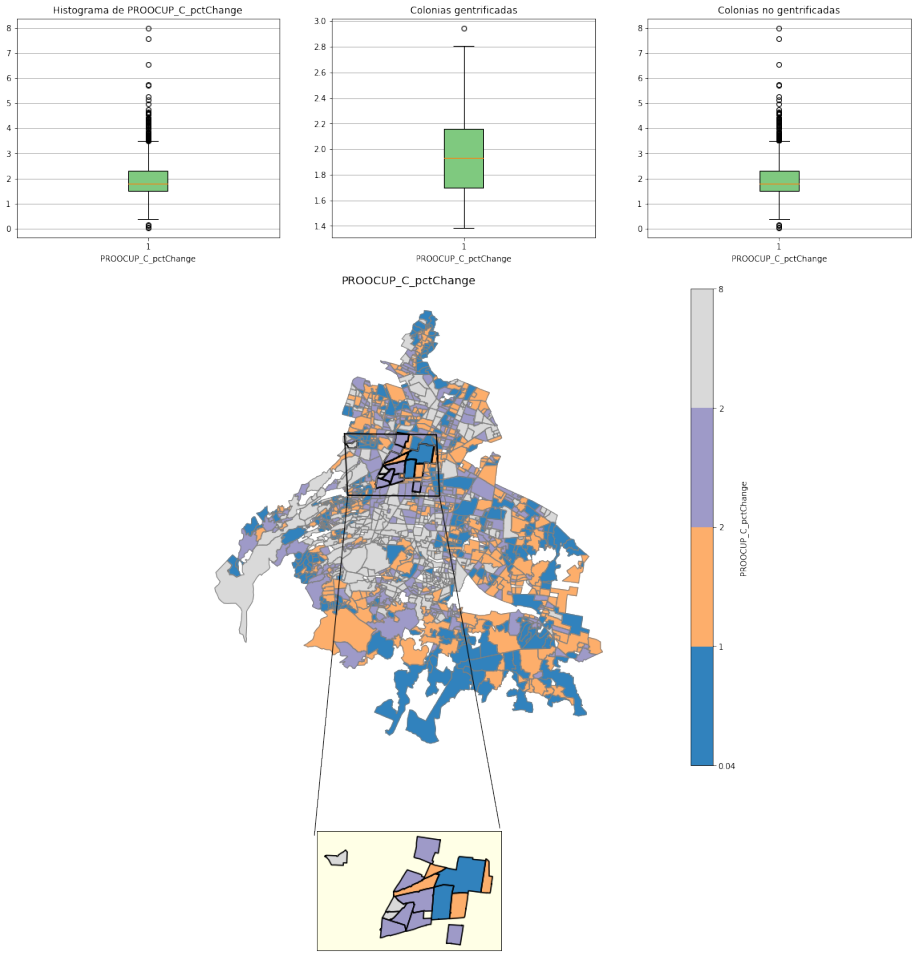

Figura 4.13: Porcentaje de cambio del promedio de ocupantes por cuarto del 2000 al 2010

La Figura 4.13 muestra el porcentaje de cambio del promedio de ocupantes por cuarto del 2000 al 2010. Se observa que la zona aledaña a Ciudad Universitaria incremento considerablemente dicho promedio debido a la presencia de estudiantes. La colonia gentrificada que presenta un cambio menor es la colonia Escandón con un incremento del 35 % del promedio de ocupantes por cuarto. Por el contrario, la colonia Hipódromo de la Condesa tienen un incremento del 280 % de dicho promedio.

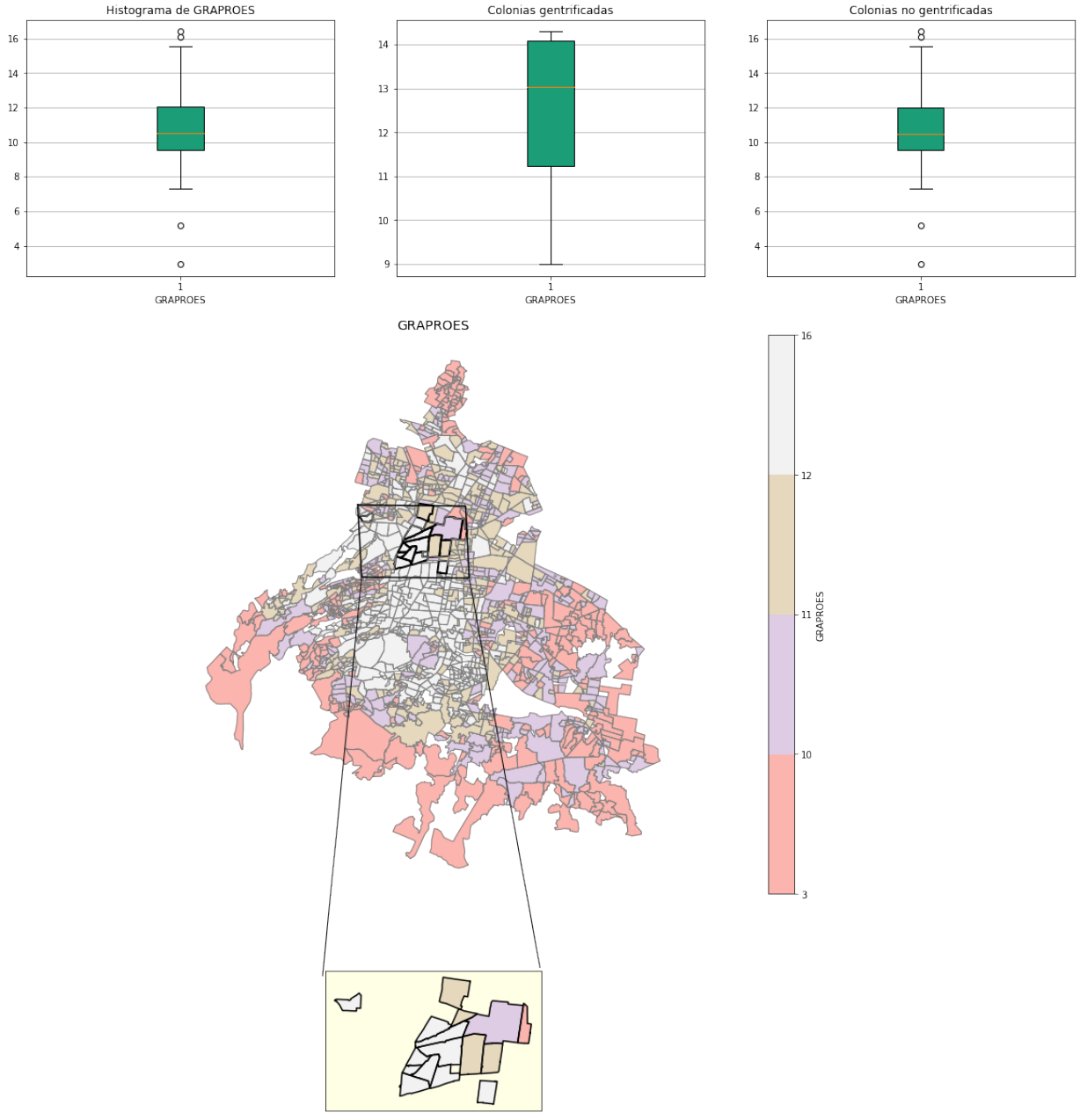

Figura 4.14: Grado promedio de escolaridad

La Figura 4.14 muestra el grado promedio de escolaridad para cada colonia. Se observa que las colonias gentrificadas tienen un grado promedio de escolaridad mayor a 9 años. La zona sur-poniente de la ciudad es la que presentan mayores valores en esta métrica.

4.2.2. Variables económicas

En esta sección se muestran las variables económicas que presentan comportamientos relevantes. En esta categoría, se tiene las variables promedio del precio del metro cuadrado para los años 2000, 2010 y 2016. Los datos originales, los cuales fueron proporcionados por la empresa Softec, son una muestra de datos que no cubre a toda la CDMX. En este caso se realiza una estimación sobre el resto de la ciudad a partir de la muestra de datos en la sección de Preparación de los datos (Ver Sección 4.21). En esta sección se muestran los mapas del cambio del precio del metro cuadrado para los años 2000, 2010 y 2016.

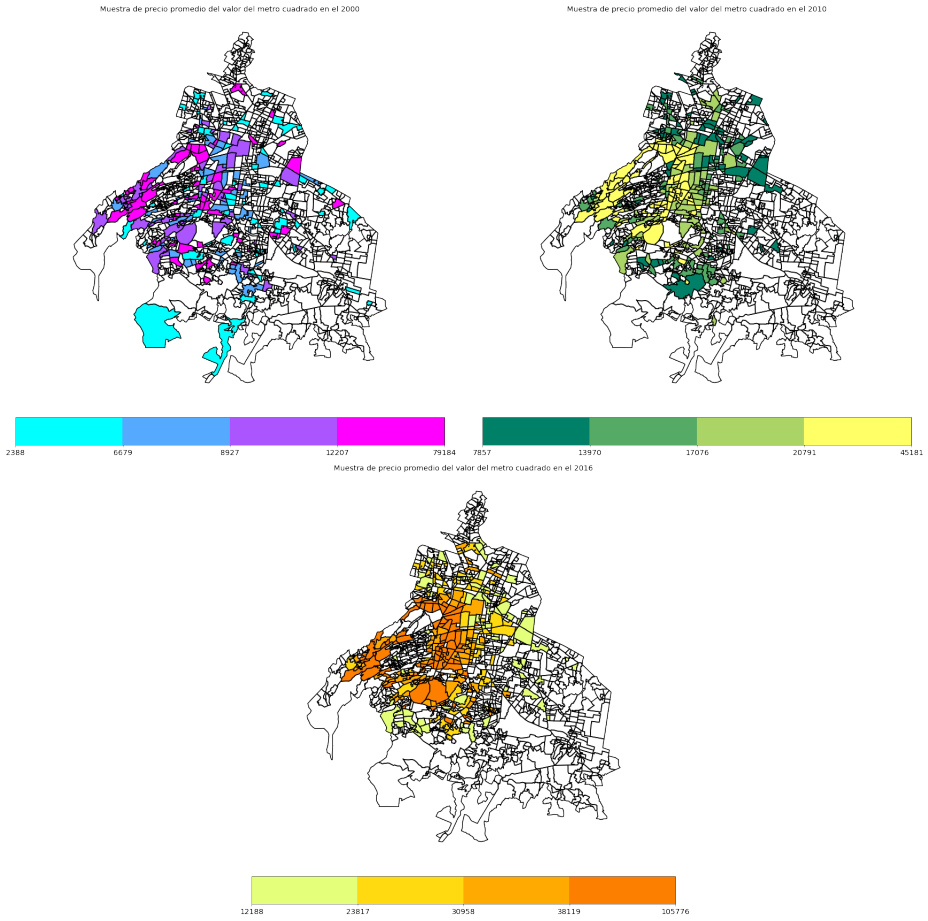

Figura 4.15: Visualización de la muestra de datos de la variable precio promedio del metro cuadrado

La Figura 4.15 muestra la visualización de la muestra de datos originales.

4.2.3. Variables de entorno urbano

En esta sección se muestra las principales variables que son consideradas en la evaluación del entorno urbano.

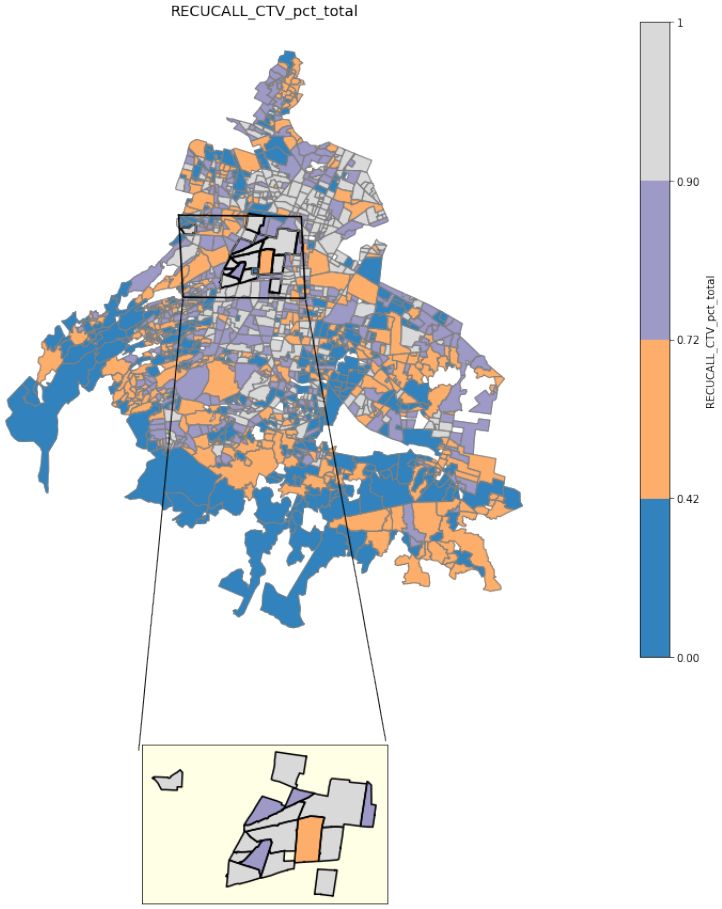

Figura 4.16: Porcentaje de vialidades con recubrimiento

En la Figura 4.16 se muestra el porcentaje de vialidades que cuentan con recubrimiento, este dato proviene del Inventario Nacional de Viviendas del 2016. Y se observa que la urbanización vial es mayor al 80 % en las colonias gentrificadas, con excepción de la colonia Doctores que tienen un 62 % de vialidades con recubrimiento. De manera general se ve mayor urbanización vial en el centro y norte de la Ciudad en comparación con la zona sur

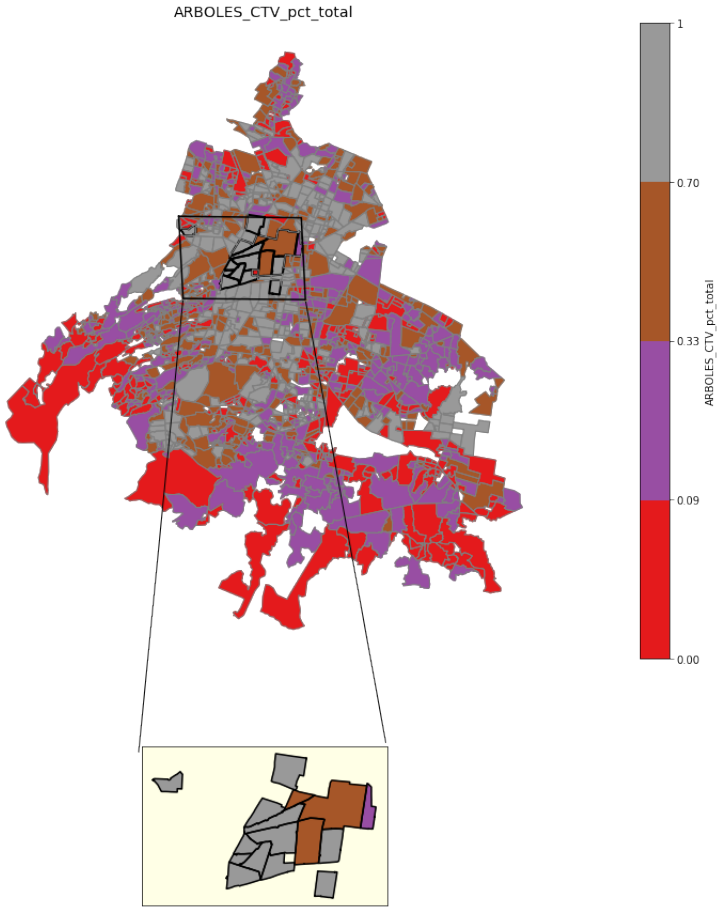

Figura 4.17: Porcentaje de vialidades que tienes arboles en todas sus vialidades

La Figura 4.17 muestra el porcentaje de vialidades que cuentan con árboles a nivel colonia. Se observa que la zona periférica sur, a pesar de ser una zona rural, cuenta con un porcentaje bajo de arboles en sus vialidades.

4.2.4. Patrones de comportamiento

En esta sección se muestra la visualización de los datos provenientes de las variables mencionadas en la Tabla 4.2. En este caso se realiza un análisis exploratorio conjunto de las variables.

Visualización T-SNE

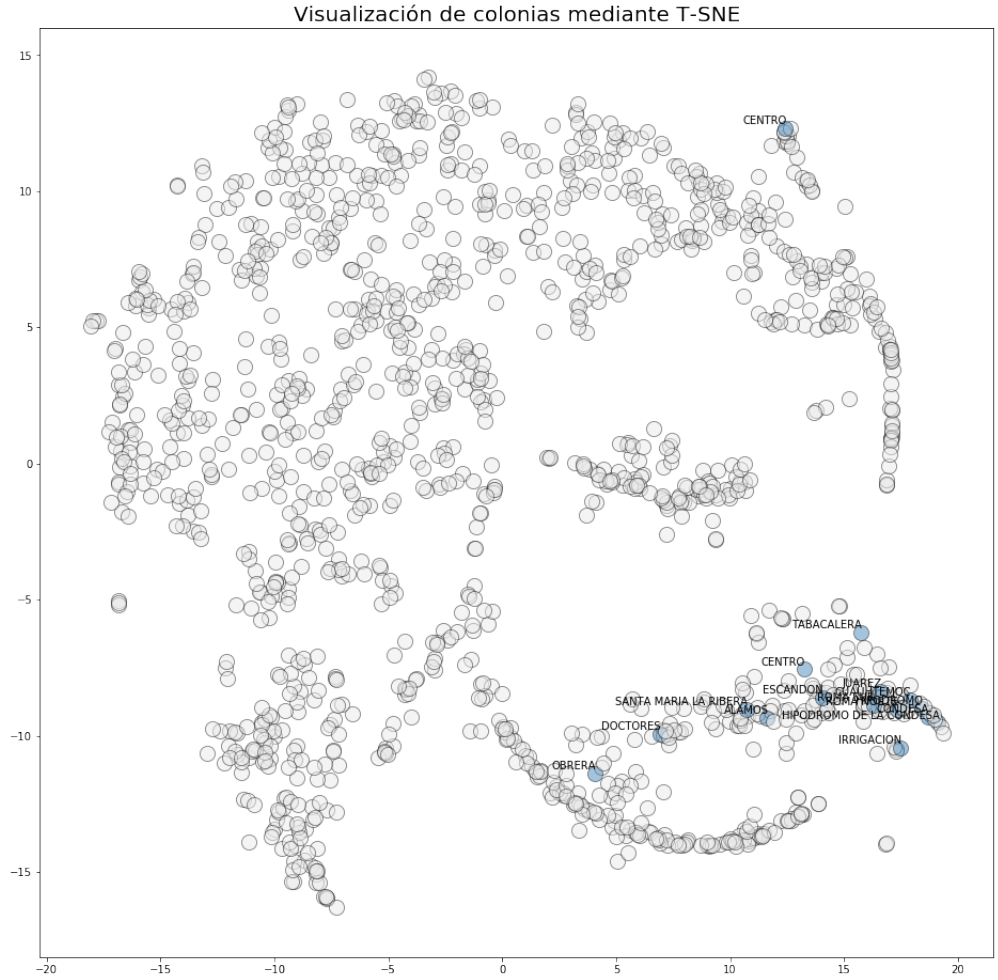

Figura 4.18: Visualización de colonias mediante la técnica de T-SNE

En la Figura 4.18 se muestra la visualización de las colonias de la CDMX usando la técnica de T-SNE (Ver Sección 2.3.1) y se observa un patrón de comportamiento de las colonias gentrificadas hacia la parte inferior derecha del gráfico. Por otro lado, coloca a las colonias restantes en la parte superior. De manera intuitiva, T-SNE está mapeando el espacio n-dimensional definido a partir de las variables a un espacio de baja dimensión que permite conservar las distancias, es decir, está colocando a las colonias gentrificadas cercanas entre sí, con excepción de la colonia Centro.

4.3. Preparación de los datos

En esta sección se realizan transformaciones de los datos para mejorar la precisión del modelo. En este caso, se completan los datos de la variable precio del metro cuadrado para los años 2000, 2010 y 2016. Se realiza una estimación bajo el ajuste de un bosque aleatorio de regresión (Ver Seccion 2.3.4).

4.3.1. Bosque Aleatorio de regresión para estimación de precios de metro cuadrado

Se tienen una muestra de datos de precios de venta de departamentos ofertados para los años 2000, 2010 y 2016, los cuales han sido proporcionados por la empresa Softec. La muestra de datos no cubre completamente a la Ciudad de México, únicamente se tiene una muestra de precios para algunas colonias. En las Figura 4.15 se muestra la cobertura geográfica de los datos muestras originales.
Se realiza una estimación del precio del metro cuadrado para cada colonia a partir de la muestra de datos. Se ajusta un bosque aleatorio de regresión (Ver Sección 2.3.4) para cada año cuya variable objetivo es el precio del metro cuadrado y las variables dependientes son los datos socio-demográficos y económicos provenientes de los Censos de Población y Vivienda realizadas por el INEGI para los años 2000 y 2010 y el Inventario Nacional de Viviendas del 2016. Es decir, se usan los datos proporcionados por el INEGI como regresores de la estimación.

La Tabla 4.2 muestra el número de datos que se tiene en la muestra de cada año, y la cantidad de variables que se usan como regresores.

Año	# datos muestra	#variables(regresores)
2000	226	228
2010	316	191
2016	334	34

Tabla 4.2: Detalle de datos disponibles para cada año

Análisis de variable objetivo

Se analiza la variable objetivo y se realiza una transformación logarítmica con la finalidad de mejorar la precisión del modelo. En la parte superior de la Figura 4.19 se muestra la distribución original de la variable objetivo para cada año y se observa que para cada año la variable se encuentra sesgada a la derecha,. En la

parte inferior de la Figura 4.19 se observa la distribución de la variable objetivo después de realizar la transformación logarítmica.

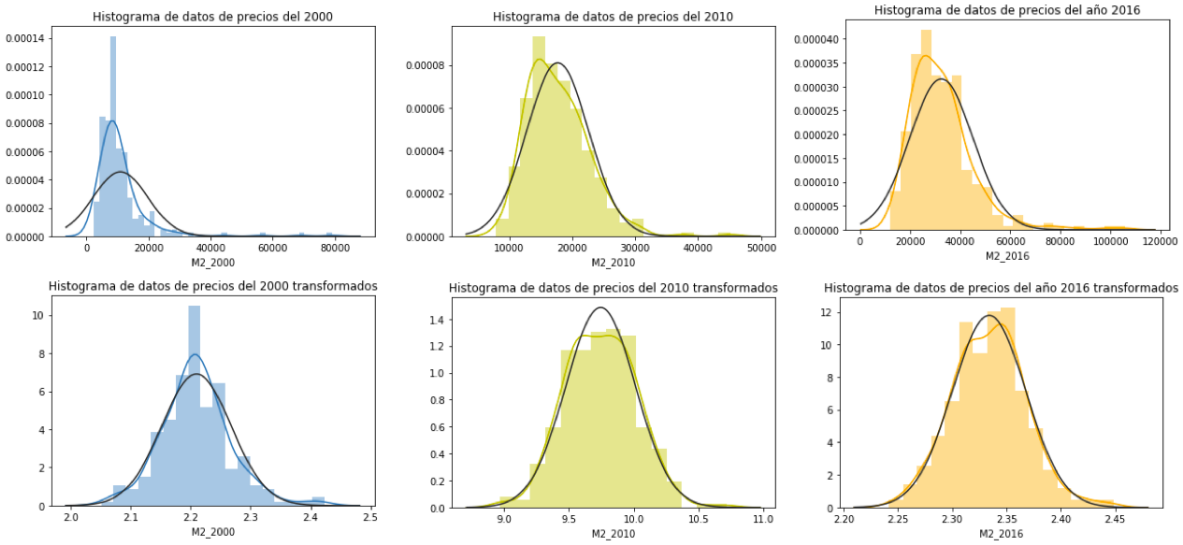

Figura 4.19: Transformación de variable objetivo

Después de transformar la variable objetivo se ajustan los modelos para cada año.

Ajuste de modelos y evaluación

Se toma como ejemplo el ajuste del modelo para el año 2016, se realiza de manera similar para el año 2000 y 2010.Se divide la muestra de datos en conjunto de prueba (*test set*) y conjunto de entrenamiento (*training set*). El *train set* se utiliza para el entrenamiento del modelo. Y el *test set* se usa para hacer estimaciones con el modelo ajustado y medir la precisión del modelo. Se utiliza el 20 % de datos para el test set y el 80 % para el *train set*.

Se ajusta un bosque aleatorio con n árboles aleatorios, donde n es el número de regresores. Para el caso del año 2016, se ajusta un bosque aleatorio con 34 árboles aleatorios. Se mide la precisión del modelo sobre el test set y se tiene que el error absoluto promedio (Mean Absolute Error, MAE) es de 0.17, y la exactitud (o *Accuracy*) del modelo sobre el test set es de 99.499 %.

En la Tabla 4.3 se muestra el resumen de métricas de precisión de modelos para los tres años. En todos los casos las métricas son calculadas para el *test set*.

Año	MAE	MSE	RSS	R^2
2000	0.46	0.39	18.084	99.99 %
2010	0.15	0.04	2.704	99.99 %
2016	0.17	0.05	3.034	99.99 %

Tabla 4.3: Métricas de precisión de ajuste de modelos para el precio del m2

Por otro lado, en la Figura 4.20 se muestra las variables más importantes para el ajuste del modelo que hace la estimación del precio del m2 para los años 2000, 2010 y 2016. Se observa que las coordenadas, promedio de ocupantes por cuarto y grado promedio de escolaridad aparecen como variables significativas en los 3 años.

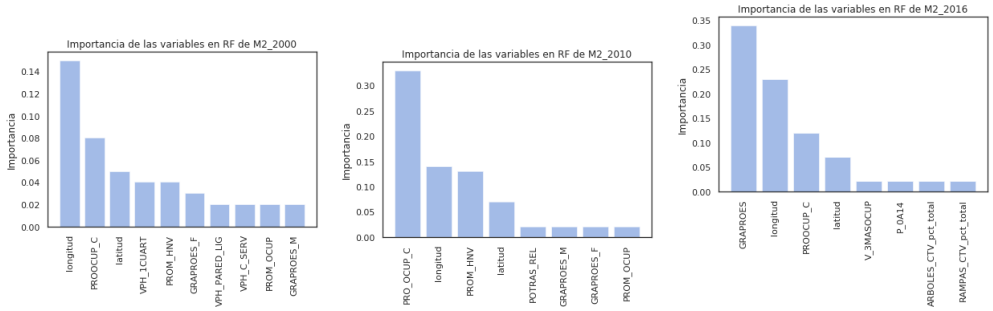

Figura 4.20: Importancia de variables para estimación de precio m2

Finalmente, después de ajustar los modelos para los tres años con una precisión mayor al 95 % se realiza la estimación sobre las colonias restantes. Los resultados de la estimación son mostrados en la Figura 4.21.

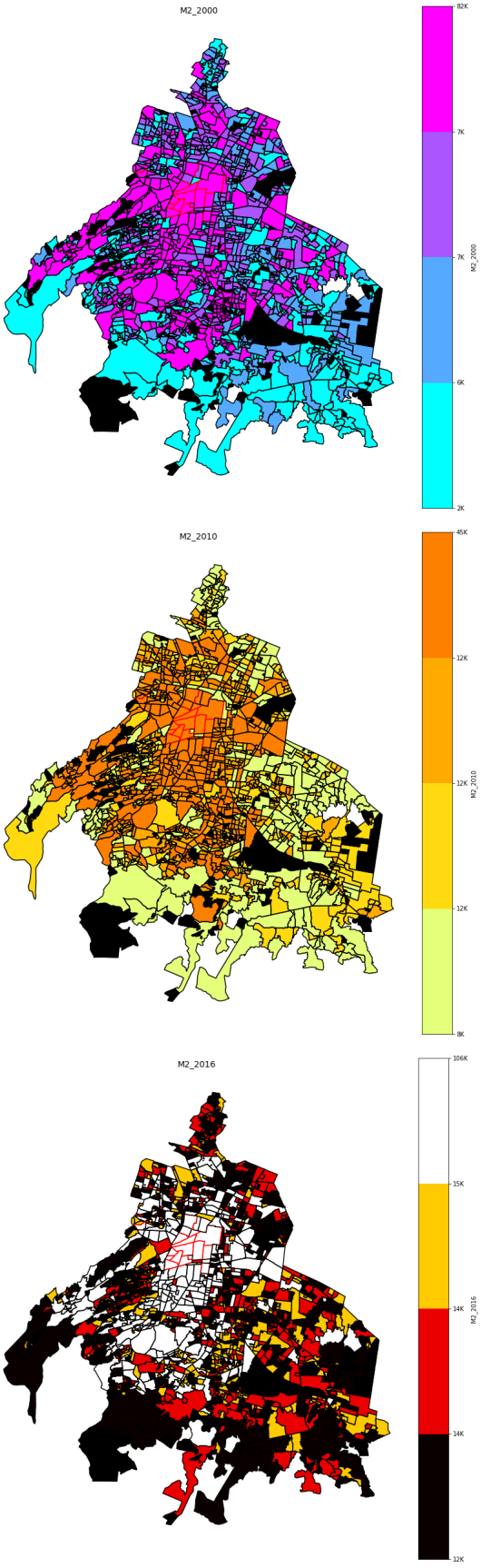

Figura 4.21: Estimación de precio del metro cuadrado en el año 2000, 2010 y 2016

La Figura 4.21 muestra la estimación del precio del metro cuadrado para cada año. En el año 2000 el precio máximo del metro cuadrado fue de $82,080 en la colonia Santa Fe Infonavit; el mínimo era de $ 2,388 en la colonia El Paraíso. Para el 2010 el máximo se incremento a $ 45,181 en la colonia Tabacalera y el mínimo a $ 7,857 en San Felipe de Jesús. Finalmente en el 2016 el precio máximo es de $ 105,776 en Chapultepec Morales y el mínimo de $ 12,188 en Santiago Acahualtepec Para los tres años la distribución de los precios altos están concentrados hacia el lado poniente de la ciudad, principalmente en las delegaciones Miguel Hidalgo, Cuahutémoc ,Benito Juárez, Coyoacán, Alvaro Obregón y Cuajimalpa, cuya zona contiene a las colonias gentrificadas.

Después de realizar dicha estimación se calcula el porcentaje de cambio del precio promedio del metro cuadrado a nivel colonia para cada periodo. Los resultados son mostrados en la Figura 4.22. Se observa que las colonias gentrificadas están contenidas en el intervalo de mayor cambio porcentual casi en su totalidad.

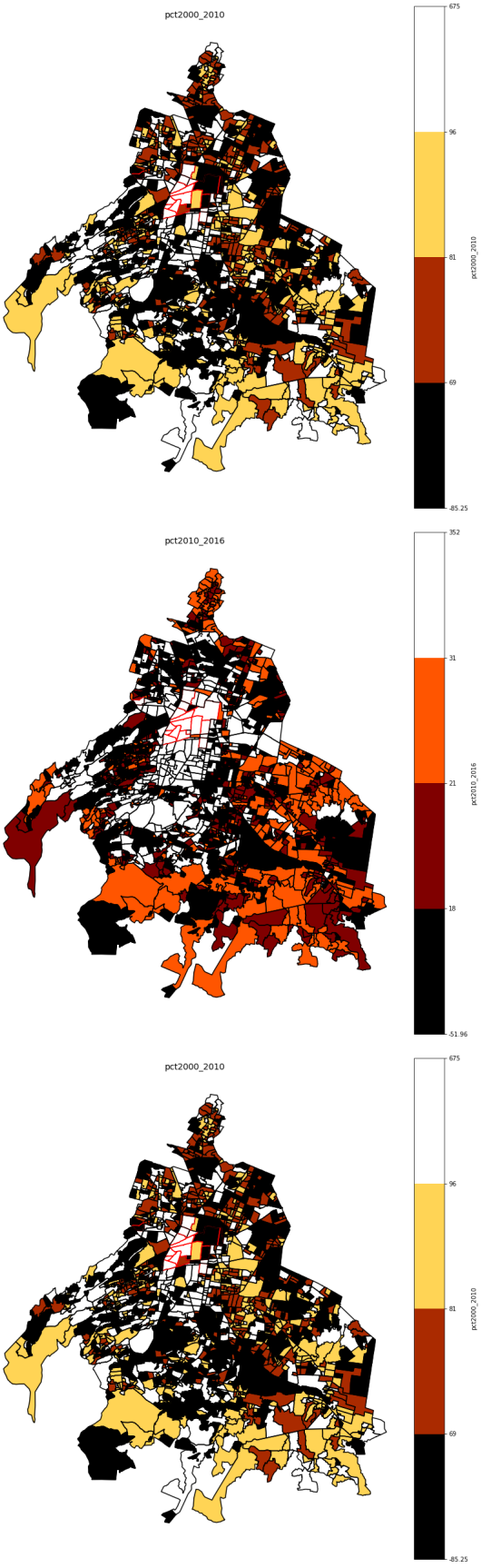

Figura 4.22: Evolución del precio de metro cuadrado del 2000 al 2016

4.4. Modelado

En esta sección se describe el proceso de modelado basado en los datos definidos en la sección anterior. Se define el tipo de problema, la variable objetivo y la evaluación del modelo. Se aborda el problema desde una enfoque de aprendizaje supervisado debido a que se desea estimar la variable objetivo "Gentrificado" y "No Gentrificado", es decir una variable categórica binaria. Se aborda la solución del problema ajustando un bosque aleatorio de clasificación (Ver Sección 2.3.4).

Otras técnicas de modelado fueron ajustados, entre ellos regresión logísticas y máquina de vectores de soporte pero el resultado de ajuste de modelos no fue satisfactorio, el modelo que muestra mejores resultado de ajuste es el bosque aleatorio de clasificación.

4.4.1. Bosque Aleatorio de clasificación

Mediante el ajuste del modelo de bosque aleatorio de clasificación se busca estimar la probabilidad de pertenencia de cada colonia de la Ciudad de México a la clase 1:"Gentrificado" o 0:"No Gentrificado"dado los valores de las variable regresoras, el cual denotaremos por X.

La expresión $Pr(Y = 1|X)$ denota la probabilidad de pertenencia a la clase 1 dado X y la expresión $Pr(Y = 0|X)$ denota la probabilidad de pertenencia a la clase 0.

En el conjunto de datos se tiene 16 observaciones con la etiqueta 1 y 1,420 observaciones con la etiqueta 0.

En el ajuste del modelo hay dos etapas a) Entrenamiento y b) Prueba. Como primer paso, se divide los datos en dos conjuntos que son usados en las dos etapas mencionadas: conjunto de entrenamiento y conjunto de prueba. El conjunto de entrenamiento se utiliza para ajustar el modelo y estimar los parámetros que permiten hacer estimaciones sobre nuevos datos. El test set se utiliza para ajustar el modelo sobre datos que el modelo nunca ha visto y se estima la precisión del modelo comparando la estimación con el valor real de la variable objetivo. La división de los datos en conjunto de entrenamiento y conjunto de prueba se hace bajo la proporción 80 % y 20 % de los datos, respectivamente, como lo muestra la Figura 4.23

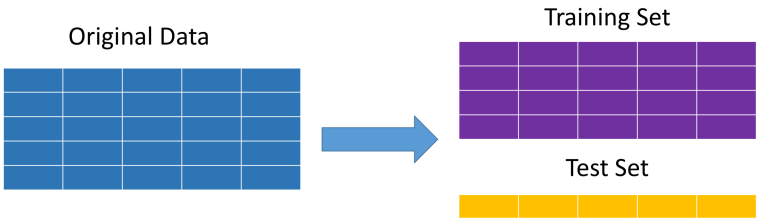

Figura 4.23: División de datos en training set y test set

Se ajusta un bosque aleatorio de clasificación con 45 árboles. A continuación de presentan las métricas que permiten medir la precisión del modelo.

4.5. Evaluación del modelo

En la Figura 4.24 se muestra la matriz de confusión (Ver Seccion 2.4). Se muestra el error de ajuste y el error de predicción.

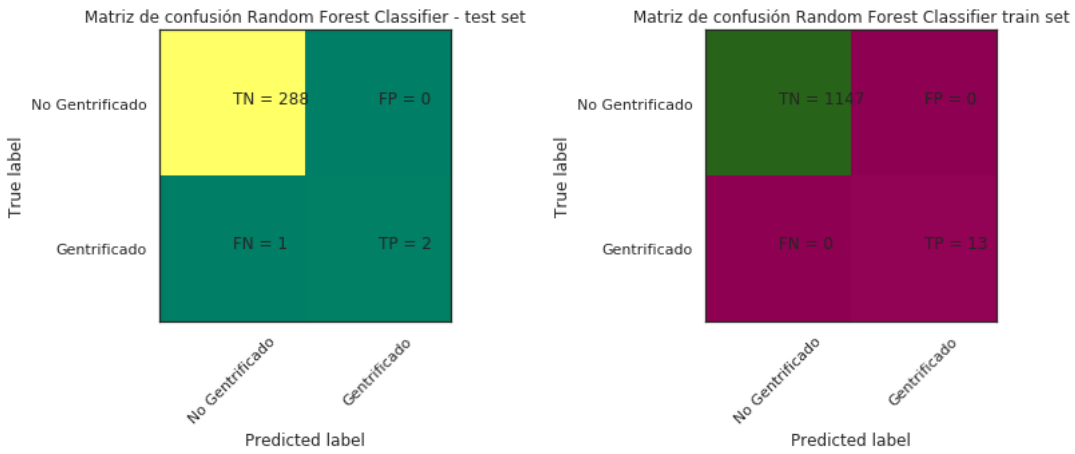

Figura 4.24: Matriz de confusión de regresión logística

En la diagonal inversa se muestran los errores del modelo. FP son los valores donde el modelo clasifica como Gentrificado a alguna colonia que no lo está y el FN sucede cuando el modelo estima como No Gentrificado a una colonia que si lo está.

Set	Clasification accuracy	Sensitivity	Precision	F1 Score
Train	100 %	100 %	100 %	100 %
Test	99.65 %	66 %	100 %	80 %

Tabla 4.4: Detalle de Métricas de precisión

La Tabla 4.4 muestra las métricas Classificacion Accuracy, Sensitivity, Precision y F1 Score. (Ver Sección 2.4).

Las métricas que importan en este caso son *Sensitivity , Precision y F1 Score*. Dado que la proporción de observaciones para cada clase no son similares, es decir, la proporción observaciones con la etiqueta 1 es menor que los que tienen etiqueta 0.

La Precisión muestra que porcentaje de colonias estimadas Gentrificadas actualmente están gentrificadas, y Sensitivity muestra que porcentaje del total observaciones fueron estimadas con la etiqueta de Gentrificado dado que son gentrificados. F1 Score es una métrica calculada a partir de Sensitivity y Precisión.

De acuerdo a estas métricas en el entrenamiento, el 100 % de las colonias estimadas como gentrificadas actualmente están gentrificadas. Y el 100 % de las colonias que están gentrificadas han sido estimadas como gentrificadas por el algoritmo.

Por otro lado, en el test el 66 % de las colonias que han sido estimadas como gentrificadas actualmente están gentrificadas. Y el 100 % de las colonias que están gentrificadas han sido estimadas como gentrificadas por el algoritmo.

En resumen, en el entrenamiento el modelo aprende identificar a una colonia gentrificada el 100 % de las veces y la clasifica como gentrificada con una precisión del 100 %. En el test el 66 % de las veces identifico a a una colonia gentrificada con una precisión del 100 %.

Por la naturaleza de este problema es más importante analizar las probabilidades de pertenencia a la clase 1 que un modelo que estime perfectamente las clases. Si el modelo estima a una colonia con una probabilidad alta de ser gentrificada es

porque encuentra patrones de comportamiento que le asignan dicha probabilidad.

Se define al grupo de colonias "gentrificables" a aquellas colonias que tienen una probabilidad mayor al 5 % de gentrificación. Analizando las probabilidades de pertenencia a la clase 1, es decir a la clase de la etiqueta Gentirifcado. En la Tabla 4.5 se muestra la lista de colonias gentrificables ordenadas por su probabilidad.

Colonia	Probabilidad	Colonia	Probabilidad
Roma Norte	0.93	Los Morales Secc Palmas	0.13
Hipodromo	0.89	San Rafael	0.13
Condesa	0.84	Lomas Altas	0.13
Santa Maria La Ribera	0.82	Granada	0.11
Hipodromo De La Condesa	0.78	Fracc Res Emperadores	0.11
Juarez	0.71	Villa Azcapotzalco	0.09
Alamos	0.71	Tacuba	0.09
Cuauhtemoc	0.71	Ciudad Universitaria	0.09
Escandon	0.67	Industrial	0.07
Doctores	0.67	Argentina Antigua	0.07
Irrigacion	0.67	Moctezuma 2Da Secc	0.07
Obrera	0.64	Agricola Oriental	0.07
Centro	0.64	Guadalupe Inn	0.07
Tabacalera	0.64	Insurgentes Mixcoac	0.07
Roma Sur	0.62	Ciudad De Los Deportes	0.07
Veronica Anzures	0.47	Algarin	0.07
Chapultepec Morales	0.42	Actipan	0.07
Del Valle	0.29	Nonoalco	0.07
Noche Buena	0.27	Las Americas	0.07
Napoles	0.24	San Miguel Chapultepec	0.07
Del Valle Sur	0.22	Pedregal Del Maurel	0.07
Polanco Reforma	0.18	Anzures	0.07
Centro	0.16	Morelos	0.07
Los Alpes	0.13	Jamaica	0.07

Tabla 4.5: Colonias gentrificables

El mapa de la Figura 4.25 se observa el mapa de probabilidad de gentrificación mostrando con color vino a las que tienen mayor probabilidad y en rosa las que tienen menos probabilidad.

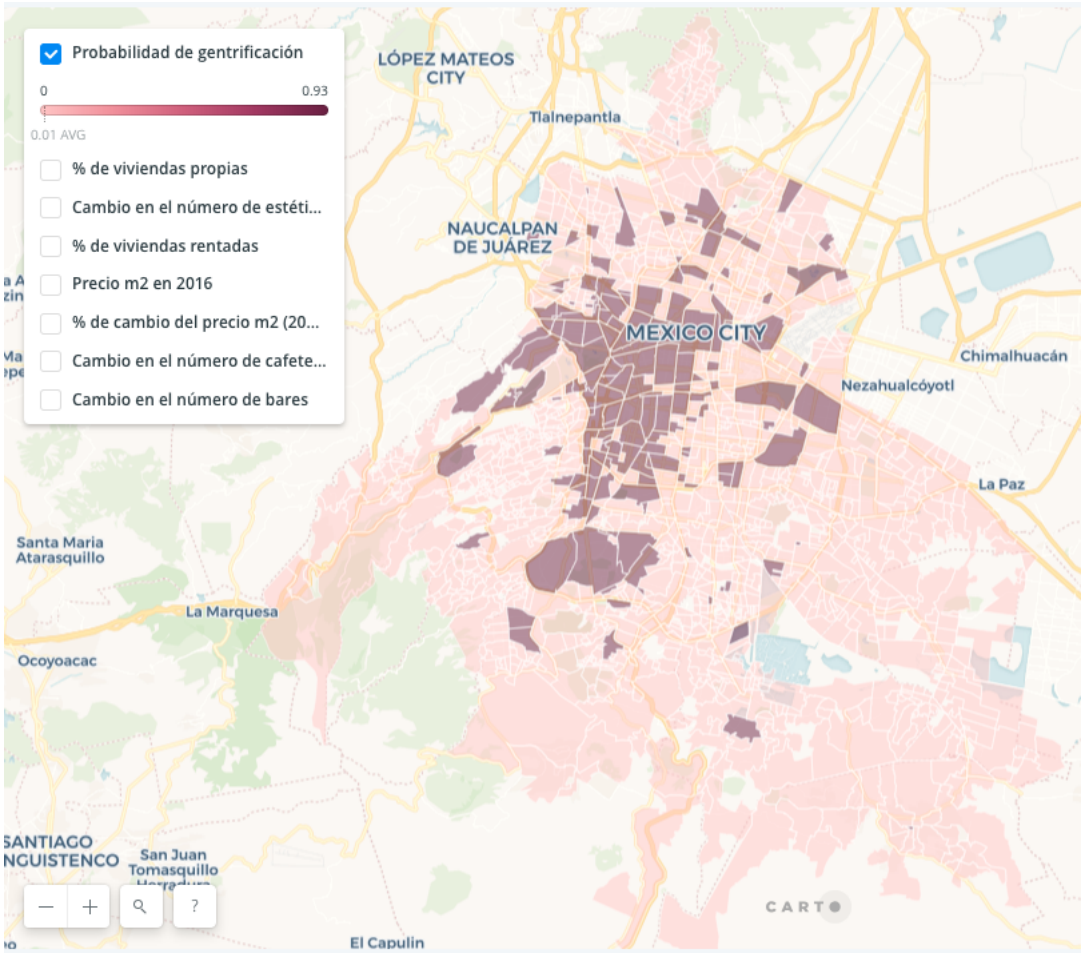

Figura 4.25: Mapa de Gentrificación

4.6. Importancia de Variables

En el ajuste del bosque aleatorio hay variables más significativas que otras. En la Figura 4.26 se observa la importancia de la variables.

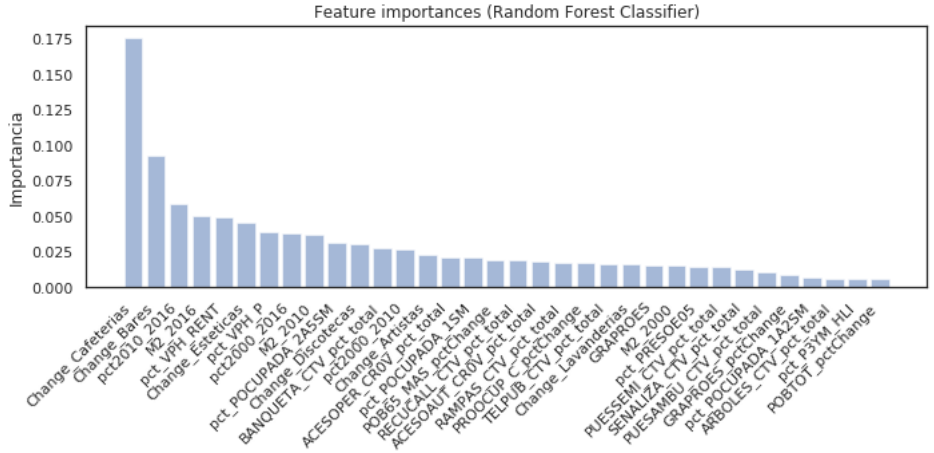

Figura 4.26: Importancia de variables

En la Figura 4.26 se observa que el número de cafeterías y bares que fueron abiertas durante el periodo 2010- 20108 y el porcentaje de cambio del precio del metro cuadrado del periodo 2000-2016 son las tres variables mas importantes para el modelo.

Capítulo 5

Conclusiones

Este capítulo contiene el resumen y conclusiones a partir del estudio realizado en el capítulo 4. Se menciona las contribuciones y futuras investigaciones.

5.1. Resumen

El problema expuesto en la sección 1.2 ha sido resuelto como se muestra en el capítulo 4. La hipótesis citada en la sección 1.3:

> *"¿Es posible calcular la probabilidad de gentrificación de las colonias de la Ciudad de México mediante el ajuste de un algoritmo de aprendizaje supervisado basado en datos socio-demográficos, económicos y de entorno urbano? "*

Dicha hipótesis tiene una respuesta afirmativa. Sí, es posible ajustar un modelo supervisado que estima la probabilidad de gentrificación de cada una de las colonias de la Ciudad de México e identificar al grupo de colonias gentrificables mediante el ajuste de un bosque aleatorio de clasificación.

La Figura 5.1 muestra dos mapas: el mapa de la izquierda se observa al grupo de colonias gentrificadas definidas en la sección 4.1, de acuerdo a las fuentes mencionadas esas colonias presentan comportamientos de gentrificación y han sido usadas para identificar el mismo patrón sobre las colonias restantes; el mapa de la derecha muestra las colonias restantes con su probabilidad de gentrificación y se observa que la periferia de la zona gentrificada inicialmente también se está gentrificando.

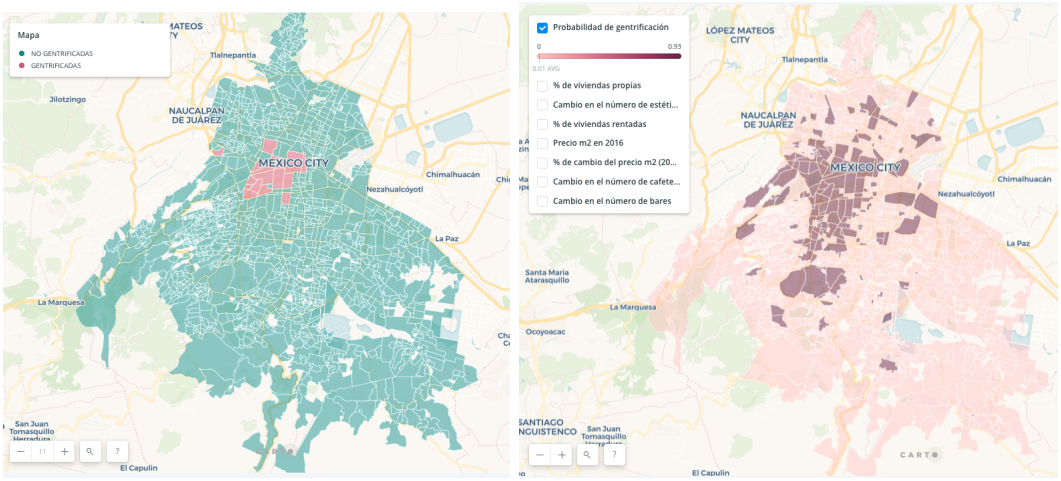

Figura 5.1: Resultados

Se observa que la zona aledaña a Ciudad Universitaria y a Unidad Profesional de Zacatenco del Instituto Politécnico Nacional (IPN) presentan significativas probabilidades de gentrificación.
Las 7 variables más significativas para el modelo son las siguientes:

- Cambio en el número de cafeterías (2010 - 2018).
- Cambio en el número de bares (2010 - 2018).
- Porcentaje de cambio del precio del metro cuadrado (2010 - 2016).
- Precio del metro cuadrado en el 2016.
- Porcentaje de viviendas rentadas.
- Cambio en el número de estéticas.

Como se menciona en la sección 2.5, las zonas gentrificadas concentran a la población con nivel socio-económico alto. En la Figura 5.2 se observa que las zonas estimadas como gentrificadas tienen altas proporciones de población que gana más de 5SM.

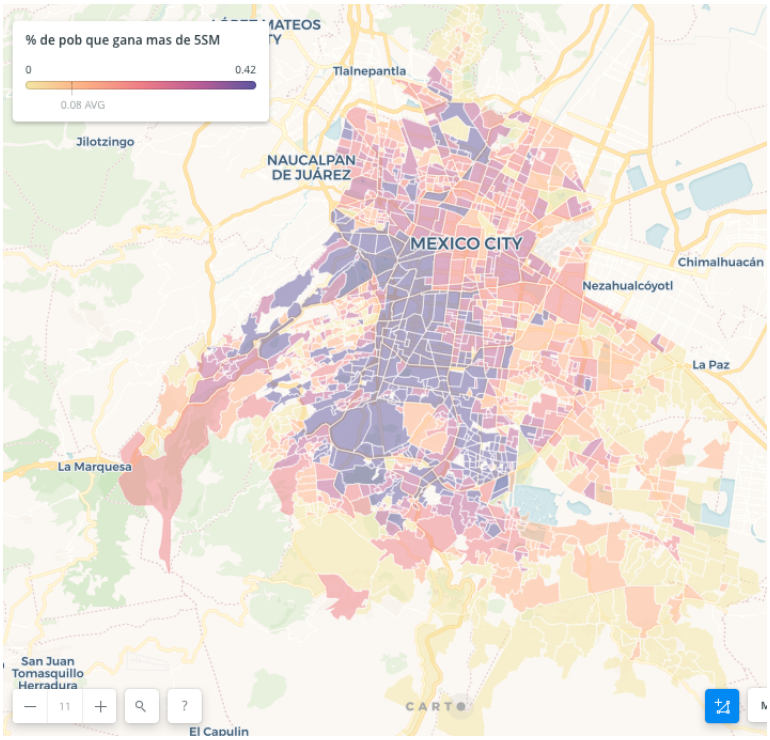

Figura 5.2: Población con alto nivel socio-económico

Sin embargo la población que reside en esta zona viven en viviendas rentadas cuyo precio de venta de m2 es elevado. Se observa que la población con nivel socio-económico bajo se concentra en la zona periférica de la ciudad, principalmente en la zona sur.

Como se menciona en la sección 2.5, la gentrificación generalmente se da por la presencia de grupos sociales. En la zona sur y norte aledañas a los principales centros de estudio UNAM e IPN, se podría estar gentrificando por la presencia de estudiantes. En la zona centro se podría estar gentrificando por la presencia de profesionistas y extranjeros con mayores niveles socio-económicos y mayor grado promedio de escolaridad que consumen servicios como los bares, cafeterías, estéticas, etc. Es por ello que estas variables son significativas para el modelo para estimar la probabilidad de gentrificación.

5.2. Contribuciones

Las contribuciones que se obtiene de este trabajo son:

- Construcción de base de datos de características socio-demográficas, económicas y de entorno urbano asociada a cada colonia de la CDMX mediante un cruce geográfico.
- Identificación de patrones de comportamiento de las colonias a partir de la visualización T-SNE.
- Implementación de modelo que permite hacer una estimación de la probabilidad de gentrificación para cada una de las colonias de la CDMX.
- Implememtación de modelo que estima el precio del metro cuadrado sobre todas las colonias de la CDMX para los años 2000, 2010 y 2016 a partir de una muestra de datos.

5.3. Trabajo futuro

El presente estudio ha estado limitado a la cobertura temporal antes definida (2000 - 2016), sin embargo podría extenderse con datos más antiguos, o bien, con datos futuros. Podría extenderse en cobertura geográfica y considerar el área metropolitana de la CDMX, así como utilizar otras técnicas de modelado.

Otra investigación interesante sería agregar datos de los grupos sociales mencionados, estudiantes y extranjeros, y medir la correlación de la presencia de estos y la influencia sobre las zonas gentrificadas.

El estudio es aplicable para otras ciudades del mundo y sería interesante estudiar la variabilidad del fenómeno de gentrificación en cada país.

Bibliografía

Agresti, A. (1997). *Statistical Methods for the Social Sciences*. Prentice Hall, 3 ed.

ANHD (2015). Anhd dap map. URL `http://map.dapmapnyc.org/`.

CAMHAJI, E. (2017). La santa que ahuyenta a los 'hipsters' del corazón de la ciudad de méxico , el país. URL `https://elpais.com`.

Cantera, S. (2017). Gentrificación: las colonias de cdmx que se "aburguesan", el universal. URL `https://www.eluniversal.com.mx`.

Chen, Z., Pears, N., Freeman, M. & Austin, J. (2011). Road vehicle classification using support vector machines. Tech. rep., Department of Computer Science University of York, York, UK.

Clark, E. (2004). The order and simplicity of gentrification: A political. *Revista Geográfica de América Central* , 256–264.

Davenport, T. H. & Prusak, L. (1998). Working knowledge: How organizations manage what they know. *Harvard Business School Press* **1**, 199.

Dulay, N. (1996). Genetic algorithms. *SURPRISE 96 Journal* **1**.

Forbes (2014). 5 colonias con potencial en el df. URL `https://www.forbes.com.mx`.

Liu, A. (2015). Data science and data scientist. *IBM Analytics* .

Loukides, M. (2010). What is data science? URL `https://www.oreilly.com/ideas/what-is-data-science`.

Maind, S. B. & Wankar, P. (2014). Research paper on basic of artificial neural network. *International Journal on Recent and Innovation Trends in Computing and Communication* **2**.

Margaritis, D. (2003). *Learning Bayesian Network Model Structure from Data*. Ph.D. thesis, School of Computer Science, Carnegie Mellon University.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

MXCity (2019). Extranjeros en la ciudad de méxico: Porcentajes cuantitativos (infográfico). URL `https://mxcity.mx`.

Olson, D. L. & Delen, D. (2008). *Advanced Data Mining Techniques*. Springer.

Ortega-Arjona, J. L. (2000). Software: Tecnología para el procesamiento de información. *Instituto de Investigaciones Antropológicas* .

Othman, A. & Sandholm, T. (2010). Decision rules and decision markets. *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems* .

Riquelme, R. (2017). Â¿en qué colonias viven los extranjeros en la cdmx? , el economista. URL `https://www.eleconomista.com.mx`.

Robert Tibshirani, D. W. T. H., Gareth James (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.

Rokach, L. & Maimon, O. (2015). *DATA MINING WITH DECISION TREES.Theory and Applications*, vol. 81. World Scientific, 2 ed.

Salinas-Arreortua, L. A. (2013). La gentrificación de la colonia condesa, ciudad de méxico. aporte para una discusión desde latinoamérica. *Revista Geográfica de América Central* **II**, 145–167.

Silwattananusarn, T. & Tuamsuk, K. (2012). Data mining and its applications for knowledge management: A literature review from 2007 to 2012. *CoRR* **abs/1210.2872**.

UCB & UCLA (2017). Urban displacement project. URL `http://www.urbandisplacement.org/`.

van der Maaten, L. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research 9* , 2579–2605.

Weiss (1999). *Introductory Statistics*. Addison Wesley.

Zhu, Y. & Xiong, Y. (2015). Defining data science. *arXiv preprint arXiv:1501.05039* .