Comparacion de Modelos de Expansion Urbana

Andrea Navarrete Rivera

2017-09-02

Contents

Chapter 1

Introduccion

La expansión urbana es un fenómeno que ha ocurrido desde siempre, en ocasiones de forma desordenada, generando diversos problemas sociales y ambientales. Por lo que para poder crear una planeación urbana se requiere entendimiento acerca de del comportamiento y los factores que generan que ciudades crezcan.

Este estudio se busca aplicar y comparar métodos de aprendizaje estadístico que aprende de los patrones del pasado, es decir de las relaciones geográficas, demográficas y económicas que han hecho crecer a la Zona Metropolitana de Chihuahua, para poder entender y poder predecir el creciemiento futuro.

1.1 Alcances

En este sentido, las contribuciones de este trabajo se pueden resumir en dos grandes objetivos: - estudiar el fenómeno de expansión urbana desde perspectivas novedosas utilizando modelos de aprendizaje estadístico. - La automatización del proceso

Chapter 2

Revision de Literatura

2.1 Expansión Urbana

La expansión de las ciudades es un fenómeno que ha ocurrido desde siempre, en ocasiones de forma desordenada, generando diversos problemas por lo que la planeación urbana cada vez requiere mayor atención. En México actualmente existen 59 zonas metropolitanas las cuales concentran el 56.8% de la población nacional (SEDESOL). Por otra parte, el crecimiento de la mancha urbana ha estado creciendo a una tasa por encima del incremento poblacional. Este crecimiento ha generado problemas de equidad, de crecimiento económico y deterioro ambiental. Debido a ello, es importante la planeación de los espacios urbanos y servicios con el fin de poder proporcionar a las personas una buena calidad de vida y construir ciudades más sustentables.

Se estima que entre 2010 y 2050 los países en desarrollo, México incluido entre ellos, incrementaran su población urbana en 2.6 billones de personas, a una tasa de 2.4% al año (United Nations Population Division 2012, file 3). Esto implica que se requerirá generar lugares residenciales para estas personas dentro de la periferia urbana para abastecer el crecimiento de la demanda de tierra para casas. A pesar de que las zonas urbanas cubran una pequeña fracción de la tierra, su expansión ha alterado de manera significativa el paisaje natural creando grande impactos en el medio ambiente y el ecosistema.

En México se estima que el 63% de la población vive en áreas urbanas y que casi la mitad se concentra en la región centro del país (Eibenshutz). El crecimiento de las áreas urbanas ha ido aumentando a una tasa por encima del incremento poblacional. Este crecimiento ha rebasado la capacidad de los tres órganos de gobierno para planear y controlar las necesidades sociales lo que ha generado problemas de equidad, ambientales y económicos. Debido a ello, es importante la planeación de servicios y espacios urbanos con el fin de poder proporcionar a las personas una buena calidad de vida y construir ciudades más sustentables. Por ejemplo, la modelación del crecimiento urbano puede crear escenarios para adaptación y mitigación enfrentando los problemas relacionados a cada ciudad.

2.1.1 Zona Metropolitana de Chihuahua

La ciudad de Chihuahua fue fundada en 1709 con categoría de Real de Minas, años más tarde adquirió la categoría de Villa con el nombre de San Felipe el Real. Durante el siglo XVIII se mantuvo con una población menor a los 6,000 habitantes con una extensión de 45 ha, en el siglo XIX su población se mantuvo por debajo de los 20,000 duplicando su extensión urbana. A finales del siglo XIX, con la llegada de la infraestructura ferroviaria se da el surgimientos de los primeros asentamientos urbanos, pero fue hasta el siglo XX cuando se da el mayor creciemiento, sumandose asentamientos aledaño y con la generación de obras públicas. (Planeación 2013)

Tabla de creciemiento

En las últimas décadas . dasdjsdlkasdad y ha crecido en los últimos años de bla a bla, y la densidad de población de bla a bla. Importancia económica que atrae a X número de personas al año. Mientras históricamente la densidad poblacional decrece, la superficie se presenta con una mayor tendencia de crecimiento Densidad de población en comparación otras zonas metropolitanas de México. Políticas de densificación

– IDEAS— El creciemento hacia el lado norte ha sido más acelerado debido a su topología plana y facilidad de dotar de servicios (Planeación 2013). Su carácter de ciudad capital del estado de Chihuahua y su impolso en el desarrollo económico ha generado una migración de las poblaciones rualres. entre 1980 y 2010, la superficie de la Zona Metropolitana de la Ciudad de México (ZMVM) creció en 257% Los nuevos desarrollos se localizan principalmente a las afueras de la ciduad.

La adecuada Planeación y Gestión de una ciudad, da como resultados un Ordenamiento Territorial armónico, equilibrado y sustentable (ejemplos y referencia)

La Ordenación Territorial de una ciudad es muy complja, ya que se compone de múltiples elementos: social, económico, político, físico y gubernamental, entre otros. De esta manera el tratar de entender los factores que y las relaciones espaciales de estos elementos es posible tener un mayor entendimiento del problema para generar acciones y decisiones más informadas. el hablar de Ordenamiento Territorial incluye el factor ambiental y está "diseñado para caracterizar, diagnosticar y proponer formas de utilización del territorio y de sus recursos naturales, bajo el enfoque de uso racional y diversificado con el acuerdo de la población." (Negrete-Bocco, 2007).

– UNir a modelos Anteriormente se han hecho trabajos sobre expansión urbana y las implicaciones por el cambio de uso de suelo con un enfoque y herramientas distintas.

2.2 Modelos de Aprendizaje Estadístico

¿Qu´e significa que un algoritmo aprenda? Grosso modo podemos decir que el aprendizaje es el proceso que convierte experiencia en conocimiento. El Aprendizaje M´aquina a su vez es un proceso mediante el cual la computadora incorpora datos sobre un fen´omeno y los convierte en experiencia; posteriormente esta experiencia se convierte en un modelo sobre el fen´omeno, que a su vez genera conocimiento sobre el mismo.

Teoría del Aprendizaje Estadístico enuncia Mitchell en su libro: "Se dice que un programa aprende de la experiencia E con respecto a una clase de tareas T y medida de desempe˜no P, si su desempe˜no sobre las tareas en T, medidas por P, aumenta con la experiencia E"

Modelo formal básico del aprendizaje Seg´un un compendio entre lo presentado por Shalev-Shwartz et al. en [55] y por Mohri et al. en [45], podemos decir que cualquier paradigma de aprendizaje requiere los siguientes elementos: Dominio.- Un conjunto Z de n individuos. Usualmente los elementos de Z se representan como vectores z = (z1, z2, .., zm) en donde zi denota una propiedad del individuo z, i = 1, …, m.

Modelo simple de generaci´on de datos.- Suponemos que los elementos de Z son generados como una muestra aleatoria por alguna densidad de probabilidad D sobre Z, de modo que son i.i.d.. Es importante mencionar que el learner (el algoritmo, la computadora) no conoce dicha distribuci´on. Este supuesto es fundamental, pues si se conociera, no habr´ıa nada que aprender.

Conjunto de entrenamiento.- S X. Es el input que el learner tiene. Se toma como subconjunto estricto porque deben quedar individuos en X que constituir´an el conjunto de prueba y /o validaci´on, que ser´a con los que midamos qu´e tan bueno es nuestro modelo. Funci´on de p´erdida.- Es una funci´on L que toma como entradas una funci´on h y a Z y devuelve un n´umero positivo: L : h × Z → R + Como podemos observar, L es una variable aleatoria. El objetivo general del Aprendizaje Estad´ıstico es minimizar la esperanza de dicha funci´on encontrando una funci´on h ∗ , i.e. min h E(L(h, Z))

didas de ´exito:
- roc
- accuracy …

Teorema (Clases de hip´otesis finitas).- PAC Learning Teorema Fundamental del Aprendizaje

Tomando en cuenta lo anterior, los algoritmos de aprendizaje se han usado exitosamente en problemas como los que menciona Mohri et al. en [45]: Detecci´on de spam Procesamiento de lenguaje natural Reconocimiento de voz y de im´agenes Biolog´ıa computacional Detecci´on de fraude Diagn´osticos m´edico Sistemas de recomendaci´on

2.2.1 Modelos no supervisados

2.2.2 Modelos supervisados

En los modelos de predicción se busca encontrar la relación existente entre las variables explicativas $X = (X^1, X^2, ..., X^p)$ y la variable de interés (y), a través de un modelo probabilístico $P(X, y)$. Esto se realiza con la finalidad de poder explicar la relación que existe entre las variables, así como para predecir futuros valores de la variable de interés, dado valores conocidos a las variables explicativas. Una forma de entenderlo, es pensar que la información generada por medio de una caja por la cual las variables explicativas X entran y por el otro lado, la variable de interés sale. Adentro de la caja hay funciones o algoritmos que relacionan las variables explicativas con la variable de interés.

Existen muchos modelos de predicción, los cuales utilizan distintas funciones para capturar la relación entre las variables. Dependerá de los datos y de la información que se conozca *a priori* para la elección de dicho modelo. A continuación se describen los modelos de predicción que se utilizarán y compararán.

2.2.2.1 Regresión Logística

El modelo de regresión logística es un modelo paramétrico de regresión que se utiliza para predecir el resultado de una variable binaria la cual toma valor de 1 si ocurre el suceso y valor de 0 si no ocurre. En este caso el suceso se referirá al hecho de que la observación pertenezca a la mancha urbana o no.

Este modelo surge a través de querer modelar las probabilidades a través de funciones lineales en x y y, al mismo tiempo de asegurarse que la suma de estas probabilidades sume uno y pertenezca al rango [0,1].Establece la siguiente relación entre la probabilidad de que ocurra el suceso a partir de un área de observación que toma ciertos valores conocidos *(x_i's)*:

$$P(y = 1|x_1, x_2, ..., x_p) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p}}$$

Realizando una transformación monotónica logística: $log[p/(1-p)]$, de esta forma tenemo una relación lineal:

$$log\frac{P(y = 0|X = x)}{P(y = 1|X = x)} = \beta_0 + \beta^T x$$

Derivado de esta relación, se estiman los parámetros $\{\beta_0, \beta_1, ..., \beta_p\}$ por medio del método de máxima verosimilitud, donde se maximiza el logaritmo de la función de verosimilitud:

$$max_\beta L(y, \beta) = \sum_{k=1}^{n} y_i ln(p_i) + (1 - y_i)ln(1 - p_i)$$

Donde n es el número de observaciones y $p_i = P(y = 1|x_1, x_2, ..., x_p)$ son las probabilidades condicionales.

Por lo tanto, para cada observación se tendrá una probabilidad estimada de que ocurra el suceso de interés, en este caso si el área de observación pertenece a la mancha urbana o no.

2.2.2.2 Árboles de Clasificación

Un árbol de clasificación es un modelo de predicción no paramétrico de aprendizaje estadístico. El método genera reglas de clasificación representadas en forma de una estructura de árbol como se muestra en la Figura. Se utilizan principalmente para hacer clasificaciones y predicciones, además asignan probabilidades, de acuerdo a la hoja a la que fueron clasificados los datos. Por otro lado, dada su estructura y metodología es posible comprender las variables que se consideran más importantes para clasificar.

En los árboles de clasificación se hacen particiones en el espacio de entradas con rectángulos con lados paralelos a los ejes. Las particiones se hacen por medio de reglas de clasificación, a partir de las cuales se busca encontrar un punto de corte el cual indique si continuar por la sub-rama derecha o por la izquierda, con la idea de hacer los nodos sucesivamente más puros, es decir, que separen de la mejor manera a nuestra variable objetivo. En la Figura, se ilustra esta idea a partir de dos variables explicativas x_1 y x_2, en donde el punto de cortes es a, a partir del cual se clasifica en a_1 o a_2.

Estos modelos no siempre resultan ser muy buenos para predecir pero son fáciles de entender y se utilizan como técnica base para otros métodos (bosques aleatorios). Por otra lado, una de sus ventajas es que son modelos robustos para valores atípicos, no es necesario transformar variables y funcionan con datos faltantes en las variables de interés. Una de las mayores desventajas de este modelo es que tiende a sobreajustar la muestra, aún y cuando existen técnicas que disminuyen este sobreajuste. Además, otra desventaja es que por la estructura que utiliza, difícilmente capturan relaciones lineales entre variables y son inestables dado que por la misma construcción pueden variar.

Algoritmo Generador

1. Se inicia con la Construcción del Árbol Maximal con la muestra de entrenamiento, el cual es un proceso recursivo:

i. El algoritmo empieza con un nodo raíz que contiene toda la muestra de entrenamiento.
ii. Para cada una de las variables explicativas, se decide la mejor forma de separar los valores de la variable objetivo, mediante una regla de partición.
iii. Se divide el nodo en cuestión en dos o más nodos hijos de acuerdo con aquella variable que mejor separa a la variable objetivo, es decir la de mayor grado de impureza.
iv. Se repite el proceso con los otros nodos hasta que no sea posible más división.
v. Por último, se elige un criterio de parada para saber cuando un nodo se declara como terminal. Estos nodos terminales son llamados hojas, los cuales contienen información sobre el número de observaciones que caen en él y la proporción para cada clase.

2. Poda del Árbol Maximal utilizando la muestra de validación.

Reglas de Particion

Las particiones de los nodos se hacen de tal manera que se reduzca la impureza del árbol, es decir, obtener mayor homogeneidad. En cada nodo se busca reducir la impureza de los nodos que le siguen, de tal forma que las hojas/regiones contengan la mayor homogeneidad posible. Las reglas de partición, dependen exclusivamente de los atributos $X = (X^1, X^2, ..., X^p)$ los cuales pueden ser tanto cuantitativos (de escala de valor y usualmente continuos) como cualitativos. Para el caso de los atributos cualitativos las reglas son de la forma:

$$\{X^j \in C\} \text{ con } C \subset \{1, ..., H\}$$

Para cada atributo cualitativo, el número de posibles reglas es:

$$\frac{2^H - 2}{2} = 2^{H-1} - 1$$

Mientras que para los atributos cuantitativos la regla se escribe de la siguiente manera:

$$\{X^j \leq c\} \cup \{X^j > c\} \text{ con } c \in \mathbb{R}$$

Asimismo, las posibles reglas serán infinitas, lo que se hace en CART es ordenar los valores que toma cada atributo X^j dentro de la muestra de entrenamiento y acotar el problema considerando sólo las reglas en las que c sea un valor intermedio entre cada par de valores consecutivos. Siguiendo ese procedimiento, el número de reglas es a lo más *n-1*, donde n es el número de observaciones de la muestra de entrenamiento. De esta manera, para ambos casos se tiene que el número de posibles reglas es finito.

Medidas de Impureza

Las medidas de impureza son funciones que sirven para determinar la elección de la regla de partición sobre todas las posibles opciones, de tal manera que se tenga una mayor homogeneidad (de la variable objetivo) en cada sub-rama. Esta medida de impureza es una función ϕ definida sobre un conjunto $(p_1, ..., p_k) \in \mathbb{R}$ de tal manera que:

$$p_j \geq 0 \forall j = 1, ...k \text{ y}$$

$$\sum_{i=1}^{k} p_j = 1$$

Además la función ϕ debe cumplir las siguientes propiedades:

- ϕ tiene un único máximo en $(\frac{1}{k}, ..., \frac{1}{k})$
- ϕ alcanza su mínimo en cero unicamente en los puntos de la forma: (1,0,…,0),(0,1,…,0),…,(0,0,…,1)
- ϕ es una función simétrica de $(p_1, ..., p_k)$

Dada una función de impureza ϕ, se puede definir la impureza para un nodo t que pertenece a un árbol T de la siguiente manera:

$$i(t) = \phi(p_1(t), ..., p_k(t))$$

donde $p_j(t)$ es la probabilidad condicional de que un elemento pertenezca a la clase j, dado que pertenece al nodo t, lo cual se estima como la proporción de elementos que caen dentro de cada clase. De este modo, la mínima impureza se obtiene cuando en un nodo t sólo hay elementos de una clase.

Para los árboles de clasificación, que tratan con variables objetivo categóricas algunas medidas de impureza son:

1. **Entropía**: es una medida utilizada en teoría de la información para medir la cantidad de información almacenada en un número de bits. En este caso una población con menor impureza tendrá un valor de entropía de 0.

$$i_{ent}(t) = \phi(p_1(t), ..., p_k(t)) = -\sum_{j=1}^{k} p_j(t) log p_j(t), \text{ definiendo } 0log0 = 0$$

2. **Índice de Gini**: esta medida es utilizada en ciencias sociales y economía; se refiere a la probabilidad de que dos cosas elegidas al azar de una población sea la misma. Una población con menor impureza tendrá un índice de Gini de 0.

$$i_{Gini}(t) = \phi(p_1(t), ..., p_k(t)) = 1 - \sum_{j=1}^{k} [p_j(t)]^2$$

   Cuando se clasifican datos sólo en dos clases no existe mucha diferencia en los resultados entre el índice Gini y Entropía, pero cuando son más clases sí. La medida de entropía normalmente tiene preferencia por grupos más pequeños y puros, mientras que el índice Gini prefiere grupos similares en tamaño.
3. **Error de Clasificación**:

$$i_{err}(t) = \phi(p_1(t), ..., p_k(t)) = 1 - max_j p_j(t)$$

4. **Prueba Ji-Cuadrada**: Es una prueba importante en estadística para medir la probabilidad de que la frecuencia observada de una muestra sea debida sólo a la variación de la muestra, es relativa a la proporción en la población original (del nodo padre), si los nodos hijos tienen la misma proporción que sus hijos, entonces el valor ji-cuadrado será cercano a cero, mientras que si sus hijos tienen menor impureza entonces el valor ji-cuadrado será alto.

Después de elegir una regla de partición en la cual se divide un nodo t en dos hijos t_1 y t_2, se define una medida de bondad para dicha partición, que en este caso llamaremos s, de la siguiente manera:

$$\Delta i(s,t) = i(t) - p_{i1}(t_1) - p_{i2}(t_2) \geq 0$$

En donde $p_{i1}(t_1) y p_{i2}(t_2)$ se refieren a la proporción de elementos del nodo t que caen en el nodo $t_1 y t_2$ correspondientes. En esta medida, es posible notar que el aumento de la bondad depende de la disminución de la impureza en los nodos hijos con relación al nodo inicial. El criterio para seleccionar la mejor partición s^* en el nodo t consiste en elegir aquella que proporciona la mayor bondad, es decir aquella que máximiza:

$$\Delta i(s^*,t) = max_s\{\Delta i(s,t)\}$$

Donde $s \in \psi$ que es el conjunto de todas las particiones posibles del nodo t.

Criterio de Parada

Asignación de Clases

Una vez construido el árbol maximal T y y los nodos terminales/hojas, el modelo le asigna a cada hoja una clase determinada. Para el caso de variables categóricas usualmente es por medio del voto mayoritario o la clase más frecuente (moda), de esta manera, se le asigna a los elementos que caen en el nodo t la clase j^* si:

$$p_{j^*}(t) = max_{i=1,..,k} p_i(t)$$

El valor $p_{j^*}(t)$ nos da el "score" dentro de cada nodo terminal, el cual es la proporción real de casos j^* en el nodo t. En los casos en donde dos clases estén empatadas en probabilidades, entonces se realiza un sorteo.

2.2.2.3 Bosques Aleatorios

El Bosque Aleatorios pertenece a los modelos de aprendizaje estadístico y surge con la idea de mejorar el desempeño de los árboles de decisión. Estos modelos surgen a partir de la idea de ensambles de modelos.

La idea general del modelo es crear distintas sub-muestras de la muestra de entrenamiento, generando árboles de clasificación con cada una y al final promediar las distintas replicaciones. De esta manera el modelo se hace más robusto que un árbol de clasificación y la varianza en los estimadores se reduce.

Reduces Overfitting It's difficult to overfit with only a subset of the available information By building the random forest model as an aggregation of weaker models, we are able to build a stronlt predective model while avoiding the pittfalls of overfitting.

Robustes Funciona bien para modelos no lineales, para captar relaciones entre parámetros y son más fácil de interpretar que otros modelos.

Bagging For some number of trees, T, and predetermined depth, D. Select a random subset of the data (convestion is roughly 2/3 with replacemenet). Train a decision tree on that data using a subset of the available features (roughly sqrt(M) by convention, where M is the total number of features). — Nathan Epstein #### Extra Árboles

2.3 Información Geo-espacial: Métodos y Proyecciones

Los modelos que se utilizan en esta tesis utilizan el Sistema de Información Geográfica (SIG) que es un sistema para la captura, almacenamiento, gestión, análisis y presentación de datos geográficos georreferenciados

(Martínez Llario 2016). De esta manera la información utilizada contiene, además de su valor, su ubicación usualmente representada en coordenadas geográficas.

La información geoespacial puede venir representada con diferente tipo de geometría: * puntos * líneas * polígonos

Una de las formas más comunes de utilizar la ubicación de los datos es por medio de las coordenadas geodesicas: latitud y longitud. Para cualquier punto en la superficie de la Tierra se puede trazar una línea recta que conecta a dicho punto con el centro de la Tierra. De esta manera la latitud del punto es el ángulo que se forma con dicha línea en la dirección norte-sur con el ecuador; mientras que la longitud es el ángulo que se forma entre dicha línea en la dirección este-oeste, con relación a un punto de partida arbitrario (generalmente el Observatorio Real en Greenwich, Inglaterra). Además, por convención los valores positivos de la latitud se encuentran en el hemisferio norte, mientras que los valores negativos en la hemisferio sur. De manera similar, los valores positivos de la longitud se encuentran al este del meridiano de Greenwich, mientras que los valores negativos al oeste.

El proyectar significa crear un plano a partir de la forma tridimensional de la Tierra, a través de una transformación matemática que busca modificar de menor manera la forma tridimensional. El matemático Carl Gausse probó que es imposible no introducir algún tipo de distorsión al generar la proyección. Existen diferentes tipos de proyecciones:

- Proyecciones Cilíndricas: Se utiliza un cilindro tangente que envuelve a la superficie de la Tierra.
- Proyecciones Cónicas: En donde se proyecta la superficie de la Tierra a un cono.
- Proyecciones Azimutales: donde se proyecta la superficie de la Tierra a una superficie plana. Estas proyecciones estan centradas alrededor de un solo punto y no muestran la superficie completa de la Tierra.

El sistema de coordenadas proyectado que se utiliza en el presente estudio es el universal transversal de Mercator (UTM), basado en la proyección cartográfica transervsa de Mercador el cual divide la superficie de la Tierra en 60 zonas secantes al meridiano. Cada zona utiliza una proyección distinta para minimizar los errores de proyección. Dentro de cada zona, las coordenadas se miden como la distancia en metros al origen de la zona, la cual es la intersección del ecuador y el centro meridiano para cada zona. La ventaja de utilizar este sistema de coordenadas se debe a la facilidad de realizar cálculos y medir distancias. Sin embargo, este sistema de coordenadas sólo es útil para áreas pequeñas que caigan dentro de una misma zona de UTM, que es el caso de nuestra área de estudio: Chihuahua.

2.3.1 Métodos Espaciales

Para poder manipular la información geo-espacial a partir de la ubicación se necesitan conocer distintos métodos.

La distancia entre dos puntos puede medirse de distintas formas:

- Distancia de coordenadas geodésicas: se utiliza la distancia del gran círculo, la cual es la longitud del semicírculo formado entre dos puntos de la superficie de la Tierra. Asumiendo que la Tierra es esférica se utiliza la fórmula de Haversine:

$$haversin\left(\frac{d}{r}\right) = haversin(\phi_2 - \phi_1) + cos(\phi_1)cos(\phi_2)haversin(\lambda_2 - \lambda_1)$$

  Donde:
  - r: es el radio de la Tierra
  - d: es la distancia euclidiana entre dos puntos
  - $\phi_1 y \phi_2$: son la latitud de cada punto respectivamente.
  - $\lambda_1 y \lambda_2$: son la longitud de cada puntos respectivamente -$haversin(\theta) = sen^2(\theta/2) = (1 - cos(\theta))/2$
- Distancia de Coordenadas Proyectadas: Esta se calcula por medio de la distancia euclidiana.

Chapter 3

Pipeline

Todo lo anterior se puede plantear de una manera específica. Utilizando métodos …

3.1 Construcción de Base de Datos

El presente estudio busca aplicar herramientas de aprendizaje estadístico para predecir la expansión urbana de la Zona Metropolitana de Chihuahua a través de características geográficas y demográficas y económicas descritas a continuación.

Debido a la naturaleza de la información de las variables se propusó crear un grid de hexagonos que cubran toda el área de influencia del modelo, en este caso la zona metropolitana de Chihuahua. De esta forma se organiza la información relevante obtenida a través de distintas fuentes dentro del grid de tal manera que cada hexagono posee la información asociada a cada área. Además a cada hexagono se le asigna una variable que indica si pertenece o no a la mancha urbana representada por los ageb urbanos.

Se decidió generar el grid con hexagonos dado a la naturaleza de las variables. Explicar porqué los hexagonos!!!!

Construcción del grid: - Grid:grid de hexagonos que cubran toda el área de influencia del modelo de 250 metros x 250 metros. - Área de influencia: Será sobre la cual se realiza la modelación y se asume que el crecimiento no superará esta área de influencia. Para Chihuahua, se consideró un radio de 7,000 metros.

sistema de coordenadas UTM de la zona 13 Norte, Datum WGS84!!!!

Herramientas utilizadas: PostGIS: extensión de la base de datos PostgreSQL que permite gestionar objetos geográficos, donde se utliza PostgreSQL (cita) como base de datos espacial en un Sistema de Información Geográfica (SIG).[1]

3.1.1 Datos demográficos

Se utilizó la información de los censos 2000 y 2010, junto con la del conteo del 2005 (INEGI 2016) a nivel AGEB.[2]

se realizó el siguiente procedimiento:

a. Se realizó una transformación de coordenadas correspondientes para tener tanto el shapefile de los AGEB como la malla dentro de las mismas unidades, en este caso UTM zona 13 Norte.

[1] Ha sido desarrollado por *Refraction Research* como un proyecto de software libre bajo licencia GLP.

[2] Área Geográfica Básica: explicar ageb

b. Se calculó el área de cada AGEB en metros.

c. Se intersectó el shapefile de AGEBs con la malla cuadriculada como se muestra en la siguiente figura. Utilizando esta intersección, se le agregó a la malla las variables elegidas del censo de población y vivienda (o conteo para 2005) de acuerdo a la proporción de área del AGEB que intersecta a cada cuadrícula.

d. Para realizarlo, primero se calculó el área de cada polígono de la intersección y se obtuvo la proporción de área que le corresponde a cada polígono de la intersección con respecto al área total del AGEB al que corresponde.

e. A partir de las proporciones de área de cada polígono con respecto al área total de AGEB que le corresponde, se calculó la proporción correspondiente para cada variable. Si la variable correspondía a una cantidad total dentro del AGEB, entonces esta variable se multiplicó por la proporción de área obtenida en el inciso (f), mientras que en el caso en el que la variable correspondiera a un promedio entonces se le asoció al polígono el dato del AGEB correspondiente.

f. A la tabla de atributos de la malla se le agregaron las variables de interés de SCINCE. Para realizar esto, se sumó o promedió, de acuerdo al tipo de variable, la información de los polígonos de la intersección que caían dentro de una misma cuadrícula del grid.

Después de haber incluido las variables de las AGEB urbanas dentro de la malla, se agregó la misma información las localidades rurales que caen dentro del área de influencia (la malla). Para esto, se utilizó la información proveniente de los principales resultados por localidad (ITER) de 2000, 2005 y 2010 (INEGI) y agregaron dentro de la malla de la siguiente manera:

i) Primero se corrigieron y homologaron las coordenadas que vienen reportadas en el ITER, debido a que estas coordenadas vienen reportadas en grados, minutos y segundos, además se corrigió el signo de la latitud. De esta manera se pudo obtener el sistema de coordenadas geográficas de las localidades [en CRS("+proj=longlat +datum=WGS84")]para generar un shapefile. Después de ser proyectado, se transformaron las coordenadas en UTM zona 13 para hacerlo coincidir con la malla.

ii) Debido a que existían diferencias en la ubicación de una misma localidad para los distintos años, esto a causa principalmente de los cambios en la precisión de los GPS, se siguen los siguientes pasos para corregir la información:

a) A las localidades rurales de 2000 y 2005 que se ubicaran en coordenadas distintas a las de 2010, se reubicaron de acuerdo a las coordenadas de 2010. De esta forma una misma localidad coincidirá en coordenadas para los tres años. El motivo de reasignarlo a las coordenadas de 2010 se debe a que la precisión de los GPS's ha aumentado; por lo que las coordenadas de 2010 tendrán una precisión mayor.

b) Siguiendo la misma lógica para las localidades rurales que en el año 2000 se encontraran en distintas coordenadas que en 2005 y además que no se encontraran en 2010, ya sea porque se unieron a una localidad urbana o simplemente desparecieron, entonces a esas las localidades de 2000 se les reasignaron sus coordenadas a las de 2005.

c) Por último, si una localidad rural se encontraba en 2000 y en 2010 pero no en 2005, entonces se generó la información de esa localidad para 2005 tomando los valores promedio de 2000 y 2010.

iii) Una vez homogeneizando las coordenadas, se incluyó la información del los las localidades rurales a la malla de la siguiente manera:

a) Se eligieron las variables de interés (las mismas de los AGEB urbanos incluidas en las mallas para 2000, 2005 y 2010). Se eliminaron las localidades con población mayor a 2,500 habitantes, puesto que estas ya son localidades urbanas y su información se encuentra incluida en los AGEB urbanos.

b) Se intersectó el shapefile de las localidades rurales con la malla (se hizo lo mismo para los tres periodos) y se le asignó la información de las localidades a las cuadrículas que intersectaran los puntos de las localidades. Si dos localidades caían dentro de una misma cuadrícula, entonces se sumaron o promediaron los valores según el tipo de variable (total o promedio). Además, si en una cuadrícula en

donde ya existía información de las AGEB urbanas intersecta una localidad rural, entonces de igual manera se suman o promedian los valores.

Utilizando la información de las localidades rurales, se incluyó a la malla la distancia mínima, en metros, del centro de cada cuadrícula a la localidad rural más cercana. Esta variable fue incluida debido a que se considera que las localidades rurales cercanas a la zona urbana tienden a migrar en busca de trabajo y servicios.

3.1.2 Económicos

Unidades Económicas

Cercanía a unidades económicas (DENUE) y al centro urbano

la distancia de cada cuadrícula a las unidades económicas grandes y medianas que se encuentran en el Directorio Estadístico Nacional de Unidades Económicas (DENUE). Para realizar esto primero se consideran unidades económicas grandes a las que registren tener a más de 250 empleados y medianas a las que registren entre 51 y 250 empleados. A partir de estas unidades se calculó la distancia mínima en metros que hay desde el centro de cada cuadrícula a estas unidades.

!! No hay datos para antes del 2008, por lo que se asume que no han cambiado, sin embargo deberían considerarse únicamente las unidades ecónomicas existentes para cada periodo.

Centro Urbano

El centro urbano de la ciudad, se caracteriza por concentrar comercios y servicios de escala urbana en la ciudad y ser sede de las oficinas de gobierno [**?**]. Por lo que se utiliza la distancia a dicho centro como una variable que influye en la expansión.

Para calcular la distancia hay centro urbano, primero se encontraron las coordenadas del centro urbano ("Google Maps" 2017) y se expandió hasta un radio de 1 km, donde se considera que se acumulan las principales actividades económicas del centro urbano. A partir de ese radio se calculo la mínima distancia en metros que hay desde el centro de cada cuadrícula al centro urbano.

3.1.3 Vías de Comunicación

Medir la accesibilidad de insumos y productos. La razón por la cual se incluye la distancia a carreteras, vías ferroviarias y aeropuerto.

Carreteras

Después se agregó la distancia que hay de cada cuadrícula a las carreteras o vías princiaples (INEGI).

La distancia se calculó como la distancia mínima que hay desde el centroide de cada cuadrícula a la red carretera o vías principales. Se consideraron las siguientes vías:

- Carreteras
- Avenidas
- Calles

Además se tomó la distancia mínima a las vías por tipo de derecho: libres o de cuota.Y por número de carriles: 1, 2 y 4 carriles

___Vías Férreas_ Las vías ferroviarias en el país son pocas — bla bla una breve descripción de cómo conectan. Eje norte-sur (Ciudad Juárez - México D.F.) y superponiente-nororiente (Cuauhtémoc - Ojinaga). ¿Quién las utiliza? Es notable el uso de un sector de la población que reside al norte de la ciudad como vialidad primaria [**?**].

3.1.4 Geográficos

Se asienta en un valle definido por serranías en los 4 puntos cardinales, así como por dos escurrimientos naturales principales o ríos. Tomando como centro el área que ocupa actualmente el centro urbano de la ciudad se tienen los siguientes límites o barreras naturales: • Al poniente la Sierra del Mogote con alturas de hasta 2000 m.s.n.m. • Al oriente la Sierra de Nombre de Dios o Cerros Colorados, más al oriente de esta, quedando en medio el valle del río Chuvíscar que actualmente acoge a la zona nororiente de la ciudad, las Sierras de San Ignacio y de Santa Eulalia • Al norte el valle de Chihuahua se abre por en medio de las Sierras del Mogote y de Nombre de Dios / Cerros Colorados • Al sur la Sierra Azul Todas las sierras referidas se consideran zonas no urbanizables debido a que cuentan con pendientes iguales o superiores al 20%.

De esta manera existe una división del espacio natural de la ciudad determinada por la extensión de la Sierra de Nombre de Dios y las elevaciones aisladas hasta casi tocar o ser parte el cerro Grande de la Sierra Azul al sur. Tanto el Cerro Grande como el Coronel son considerados por el PDU 2000 como zonas con potencial para decretarse patrimonio natural, e inclusive forman parte del escudo de armas de la capital. (IMPLAN)

La pendiente se construyó a partir de la altura de los Modelos Digitales de Elevación (INEGI 2000 a escala de 1:250 metros) la cual viene en formato raster.

Se extrajo la información del raster para incluirla a la malla por medio una función que se aplica a los valores del raster (pixeles) que caen dentro de una misma cuadrícula. Se crearon tres varibles distintas una para cada función: - slope_mean: utilizando el promedio
- slope_med: utilizando la mediana
- slope_max: utilizando el máximo

Después de las variables económicas se agregó una variable que indica el tipo de uso de suelo al que pertenece cada cuadrícula, de acuerdo con las Capas de Uso de Suelo III, IV y V de INEGI(a escala 1:250 000) para cada una de las mallas del año que le corresponda. La información de las capas de uso de suelo muestra la distribución geográfica de los diferentes tipos de vegetación natural (primaria y secundaria) y de la vegetación inducida. La variable se agregó de la siguiente manera:

i) Con la información de INEGi de las capas de uso de suelo:

- Serie III de 2002-2005 con año de referencia 2002 que utiliza imágenes LANDSATTM (30m)
- Serie IV de 2006-2010 con año de referencia 2007 que utiliza imágenes SPOT 5 (10m).
- Serie V de 2011- 2013 con año de referencia 2011 que utiliza imágenes LANDSAT (30m)

ii) Las categorías de las capas de uso de suelo de las tres series, se reclasificaron y agruparon en 10 categorías, mostradas en la tabla siguiente:

3.1.5 Label

Con el fin de poder medir las causas de expansión urbana, se define en este trabajo la variable de interés (a predecir) como un cambio en el uso de suelo de no-urbano a urbano, en un periodo de tiempo determinado.

3.2 Modelos

Al realizar los modelos predictivos, se tomararán tres conjuntos de datos:

- Muestra de Entrenamiento: en este caso son los datos de 2000 (t_1) para predecir la mancha urbana de 2005 (t_2). Estos datos son los que se utilizan para entrenar los modelos.

$$t_1 \rightarrow t_2$$

- Muestra de Prueba: en este caso se utilizan los datos de 2005 (t_2) para predecir la mancha urbana de 2010 (t_3).Con esta muestra es posible comparar el desempeño de los diferentes modelos. Estos mismo datos nos darán el error real cometido con cada modelo. Esta comparación se realiza analizando qué tanto se ajustan los valores predichos a los reales.

$$t_2 \rightarrow t_3$$

- Muestra de Validación: son los datos que se utilizan para predecir 2015 (t_4).

$$t_3 \rightarrow t_4$$

Se corrieron los modelos: - Regresión Lineal (grid parámetros) - Bosque Aleatorio (grid parámetros) - Extra Árboles (grid parámetros) -

Comparar con lo que se usa normalmente

Chapter 4

Resultados

Lees de postgres los resultados los mejores 5 -> graficas en la métrica que quiera este es el que voy a utlizar para predecir 2015 ah le atine! si fracaso- > esto muestra que hay que tener mayor consideración de los datos

Chapter 5

Conclusiones

automata celular Capitulo 1 con resultados

Trabajo futuro agregar: - Las Líneas de alta tension - los acueductos

"Google Maps." 2017. "https://www.google.com.mx/maps/place/Palacio+de+Gobierno/@28.6391952,-106.0754889,17z/data=!3m1!4b1!4m5!3m4!1s0x86ea4351999c130b:0x6f6ddd8d93dc161!8m2!3d28.6391905!4d-106.0732948".

INEGI. 2016. "Insitituto Nacional de Estadística Y Geografía." "https://www.beta.inegi.org.mx/".

Martínez Llario, José C. 2016. *PostGis 2, Análisis Espacial Avanzado*. cartosig.upv.es.

Planeación, IMPLAN: Instituto Municipal de. 2013. "Plan de Desarrollo Urbano 2040." "http://www.implanchihuahua.gob.m 2040/psmus/DIAGNOSTICO_URBANO. pdf".