Detección de fraude fiscal usando análisis no supervisado

http://www.alianzacaoba.co

Outline

Introducción

Descripción del Problema

- En Bogotá, la secretaría de hacienda está encargada del recaudo de ciertos impuestos.
- Por ejemplo, está encargada del recaudo del impuesto de delineación urbana.
- El impuestos de delineación urbana se paga al realizar una obra nueva, una remodelación, una ampliación, etc.

Descripción del Problema

- Un contribuyente declara el costo de la obra, y pagan en impuestos un porcentaje de este valor.
- En el caso de delineación, el impuesto a pagar es el 2,6 % del costo declarado.
- En general, los impuestos se pagan sobre un porcentaje de un valor declarado.

Motivación

Motivación

- La base para el cobro de impuestos es una cantidad dada por el contribuyente.
- El contribuyente no tiene ningún incentivo para declarar un ingreso o un costo verdadero.
- La consistencia del pago de ciertos impuestos es más fácil de verificar para algunos impuestos.

- En la literatura, se suele usar análisis supervisado en datasets pequeños de auditorías.
- También se suelen hacer ejercicios usando series de tiempo y desviaciones históricas.

Motivación

- No tenemos datos históricos sobre los contribuyentes.
- No tenemos datos marcados sobre obras auditadas y marcadas como fraudulentas.
- ¿Qué podemos hacer?

Metolodología

- Para poder atacar el problema concreto de fraude en el impuesto de delineación urbana, nos basamos en una premisa fundamental.
- Obras parecidas deberían pagar cantidades de impuestos parecidas.

Metolodología

- Basados en la anterior premisa, proponemos una metodología de análisis no supervisado.
- Primero, separamos las obras en grupos, donde dos obras pertenecen al mismo grupo si son similares.
- Luego, ajustamos una distribución de probabilidad a los costos de cada obra para cada grupo.
- Tomamos como anómalas las obras cuyo costo es menor al primer decil de la distribución de su grupo.

Clustering espectral

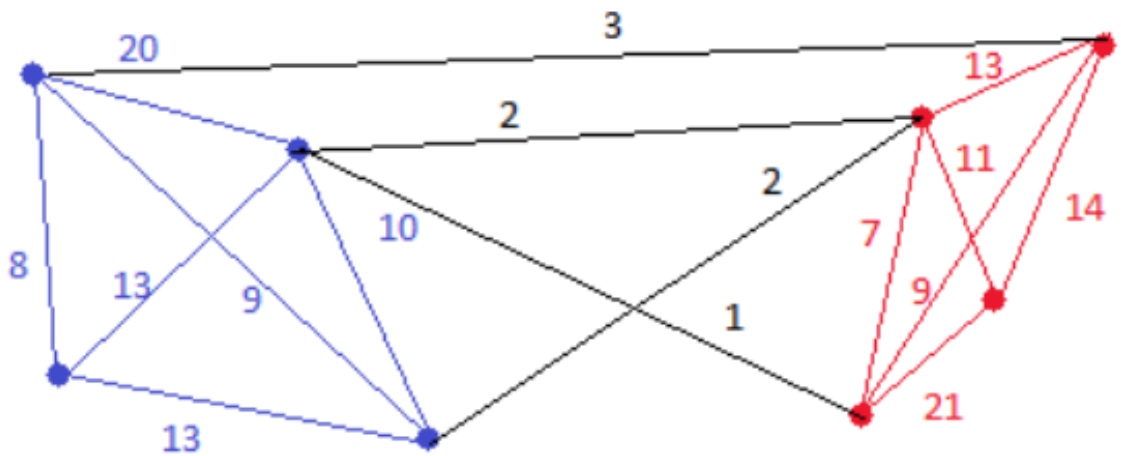
- Para separar las obras en grupos, vamos a utilizar variables que las describen.
- Estas variables son, por ejemplo, duración de la obra, estrato, UPZ, metros construidos, número de pisos habitables, área del lote, etc.
- Así, podemos pensar en las obras como puntos en $\mathbb{R}^n + 1$ y usar un algoritmo de clustering para agruparlas. Por ejemplo, **K-medias**.

K-medias

- el algoritmo más conocido para la construcción de grupos, **K-medias**, tiene varios defectos.
- No se pueden usar variables categóricas.
- Solo separa regiones convexas.
- Una alternativa son las mixturas gaussianas, pero estas asumen que las variables que describen los objetos siguen una distribución dada.

Clustering espectral

- Parar poder abordar el problema, vamos a reformularlo a uno de detección de comunidades en grafos pesados.



Clustering espectral

- Una comunidad es un conjunto de nodos cuyas aristas entre ellos tienen peso grande y cuyas aristas con nodos afuera de la comunidad tienen peso bajo.
- Una comunidad es maximal si no podemos agregarle más nodos.
- Notemos que esta no es una definición formal.

Detección de cliques

- Este problema se suele conocer como el *problema de detección de cliques*.
- Este problema tiene varias versiones, por ejemplo el problema de **decisión de cliques**: decidir si en un grafo existe un clique de tamaño mayor a un número dado.
- El problema de decisión de cliques es NP-completo.

Clustering espectral

- un algoritmo bastante conocido para la detección de cliques es *spectral clustering*.
- La entrada de este algoritmo es un grafo pesado, y el número k de cliques que se van a construir.
- Procederemos a explicar cómo funciona este algoritmo.

clustering espectral

La idea fundamental del spectral clustering es la siguiente:

1. Transformar los puntos que queremos agrupar de manera no lineal para que sea más fácil detectar su estructura de grupos.
2. Usar un algoritmo tradicional de clustering para agruparlos.

Clustering espectral

1. Para un grafo G Sea D su matriz de adjacencia.
2. Definimos A como la matriz de similaridad de G.

$$A_{ij} = e^{-\frac{d(X_i, X_j)}{2\sigma^2}}$$

- σ es un parámetro de regularización.
- A es una matriz que mide que tan similares son los nodos del grafo.

Clustering espectral

- Para un vértice i en G, definimos su peso p_i como

$$p_i = \sum_j A_{ij}$$

- Definimos la matriz de peso de G como la matriz D cuyas entradas en la diagonal son los pesos p_i.
- Definimos la matriz laplaciana del grafo como:

$$L := D - A$$

Matriz laplaciana

La matriz laplaciana tiene muchas propiedades importantes .(Spectral graph theory)

- L es simétrica.
- L es semidefinida positiva (y por lo tanto todos sus valores propios son no-negativos)
- El valor propio más pequeño de L es 0 (asociado al vector propio 1).
- La multiplicidad del vector propio 1 corresponde al número de componentes conectadas de G.

Ratio cut

- Para dos subconjuntos de vértices V y U consideramos la función

$$cut(U, V) = \sum_{U \in A,\ j \in V} w_{ij}$$

donde w_{ij} es la similaridad entre los vértices i y j.

- Encontrar una partición óptima del grafo en subgrafos máximales corresponde a solucionar el problema

$$minCut(U_1, .., U_k) = min \sum_{i}^{k} cut(U_i, U_i^c).$$

Ratio cut

- Este problema es fácil de solucionar, el problema es que suele contruir clusters de 1 solo vértice.
- Para evitar eso, controlamos por el tamaño de los clusters.
- Definimos el problema ratioCut como:

$$minCut(U_1, .., U_k) = min \sum_{i}^{k} \frac{cut(U_i, U_i^c)}{|U_i|} cut(U, U^c).$$

Ratio cut

- El problema ratioCut es NP.
- Dada una partición de G en subgrafos $U_1, ..U_k$,definimos una matriz $H \in \mathbb{R}^{n \times k}$ como:

$$H_{ij} = \begin{cases} \frac{1}{\sqrt{|U_i|}} \; si \; i \in A_j. \\ 0 \text{ de lo contrario.} \end{cases}$$

- Se puede mostrar (fácilmente) que solucionar Ratio cut es equivalente a solucionar el problema:

$$\min_{U_1,..,U_k} Tr(H^T LH) \text{ sujeto a } H^T H = I.$$

- Para solucionar más fácilmente el problema, lo relajamos, para obtener:

$$\min_{\mathbb{R}^{n\times k}} Tr(H^T LH) \text{ sujeto a } H^T H = I.$$

- El teorema de Rayleigh-Ritz muestra que la solución $\hat{H}$ de este problema es la matriz que contiene los k primeros valores propios de L como columnas.

El algoritmo

1. Entradas: G y k
2. Construya L.
3. Encuentre los k vectores propios más pequeños de L (menos 0).
4. Construya la matriz H cuyas columnas son los vectores propios y_i encontrados.
5. Utilice k-medias para hacer clusters $C_1, .., C_k$ con los vectores y_i.
6. Asigne a un vértice x_i el cluster C_j si $y \in C_j$.

Construcción de comunidades de obras

- Vamos a construir el grafo de obras.
- Primero, definimos una **métrica** entre las obras: Dada una obra $X_i \in \mathbb{R}^{n+1}$ y $j \in \{1, .., n\}$, X_i^j denota la j-ésima cordenada de X_i.

Construcción de comunidades de obras

- Dadas dos observarciones X_k, X_l definimos la j-ésima distancia d_j como:

$$d_j(x_k, x_l) = \begin{cases} 1 & \text{Si j es una variable categórica y} \\ & x_k^j, x_l^j \text{pertenecen a la misma clase.} \\ 0 & \text{Si j es una variable categórica y} \\ & x_k^j, x_l^j \text{ no pertenecen a la misma clase.} \end{cases} \quad (1)$$

$$d_j(x_k, x_l) = \frac{(x_k^j - x_l^j)^2}{1 + (x_k^j - x_l^j)^2} \text{si j es una variable continua.} \quad (2)$$

Construcción de comunidades de obras

- Notemos que d_j es una métrica para cada j.
- Finalmente, definimos la distancia entre X_l y X_k como:

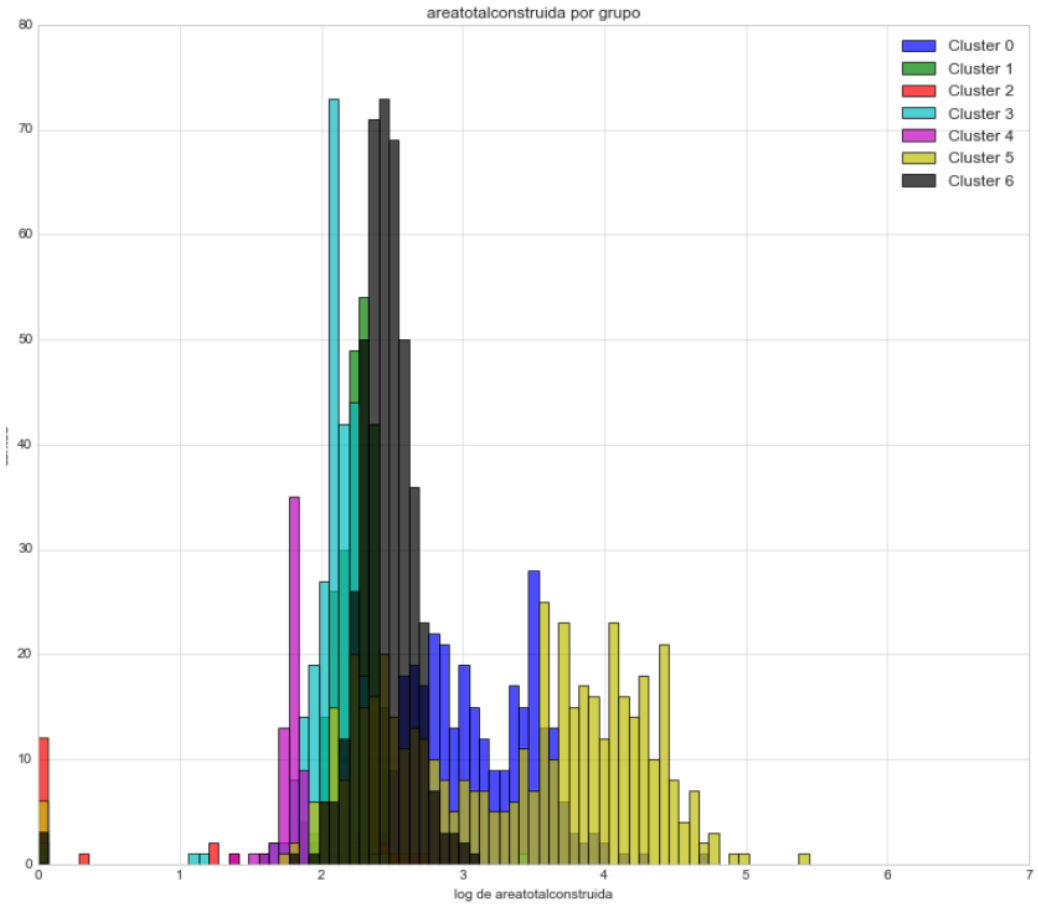
$$d(X_k, X_l) = \sum_{j=1}^{n} d_j(X_k, X_l)$$

- Como d es una suma de métricas, es una métrica.
- Construimos el grafo cuyos nodos son obras y el peso de las aristas es la distancia entre dos obras.

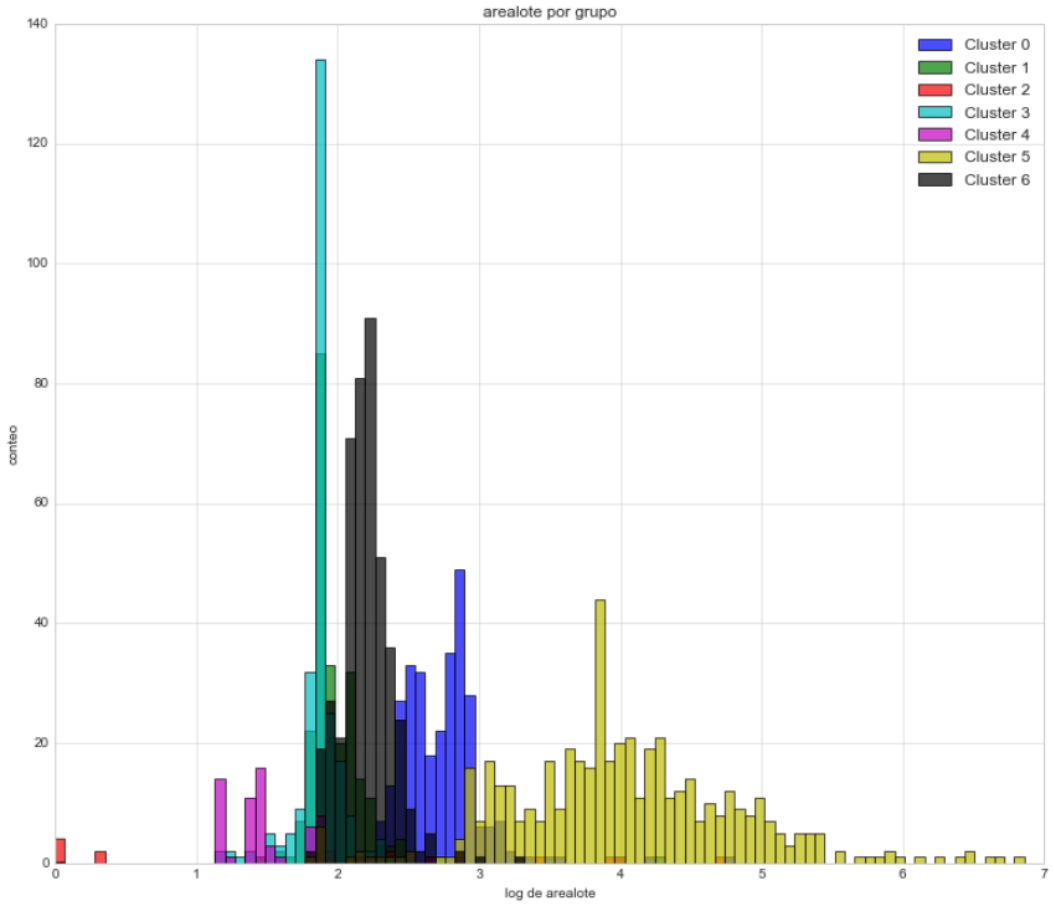
Construcción de comunidades de obras

- Usamos spectral clustering para construir las comunidades.
- Corremos de nuevo el algoritmo en las comunidades más grandes (hierarchical clustering).
- Usando un dataset de 1800 obras, obtenemos 7 comunidades.

Separación de variables



Separación de variables



Detección de fraude fiscal

Estimación por kernels

- Asignamos una distribución de probabilidad al costo de las obras dentro de cada comunidad construida.
- Para esto, usamos **Estimación por kernels**.
- La estimación por kernels es una técnica de estadística no paramétrica que permite aproximar la distribución subyacente de un conjunto de datos i.i.d.

Estimación por kernels

Supongamos que tenemos un conjunto de observaciones $x_1, .., x_p$ independientes, muestreados de una distribución desconocida f. El estimador por kernels de f es:

$$\hat{f}_h(x) = \frac{1}{ph} \sum_{i=1}^{p} K(\frac{x - x_i}{h}).$$

Estimación por kernels

1. $K()$ es kernel: es una función no negativa cuya integral es 1. En general, como en este caso, se usa la distribución gaussiana.
2. h es un parámetro de regularización que pesa la influencia de cada una de las observaciones dentro de la distribución final. Se conoce como el ancho de banda.
3. p es el número de observaciones.

Estimación por kernels

- Intuitivamente, la estimación por kernels asigna a cada punto una pequeña curva, para indicar que es más probable observar puntos cerca a este punto, que observar puntos alejados.
- Se suman todas estas curvas.
- Se normaliza para obtener efectivamente una distribución de probabilidad.

Estimación por kernels

- Bajo condiciones poco restrictivas, el estimador por kernels es buen estimador de la distribución real, en el sentido que:

$$E[(\hat{f}_p^h(x) - f(x))^2] \to 0 \text{ cuando } p \to \infty.$$

Detección de fraude

- Para una obra X sea X^{n+1} el costo de la obra.
- Para una comunidad $q \in \{1, .., m\}$ sean $X_{q1}, .., X_{ql}$ las obras que pertenecen a esta comunidad.
- Estimamos la distribución de los costos declarados $\hat{f}_q$ usando estimación por kernels en los valoresX_{qi}^{p+1}, $i \in \{1, .., q_l\}$:

$$\hat{f}_q := \frac{1}{q_l} \sum_{i=1}^{q_l} K(\frac{x - X_{iq}^{p+1}}{h})$$

Distribuciones estimadas

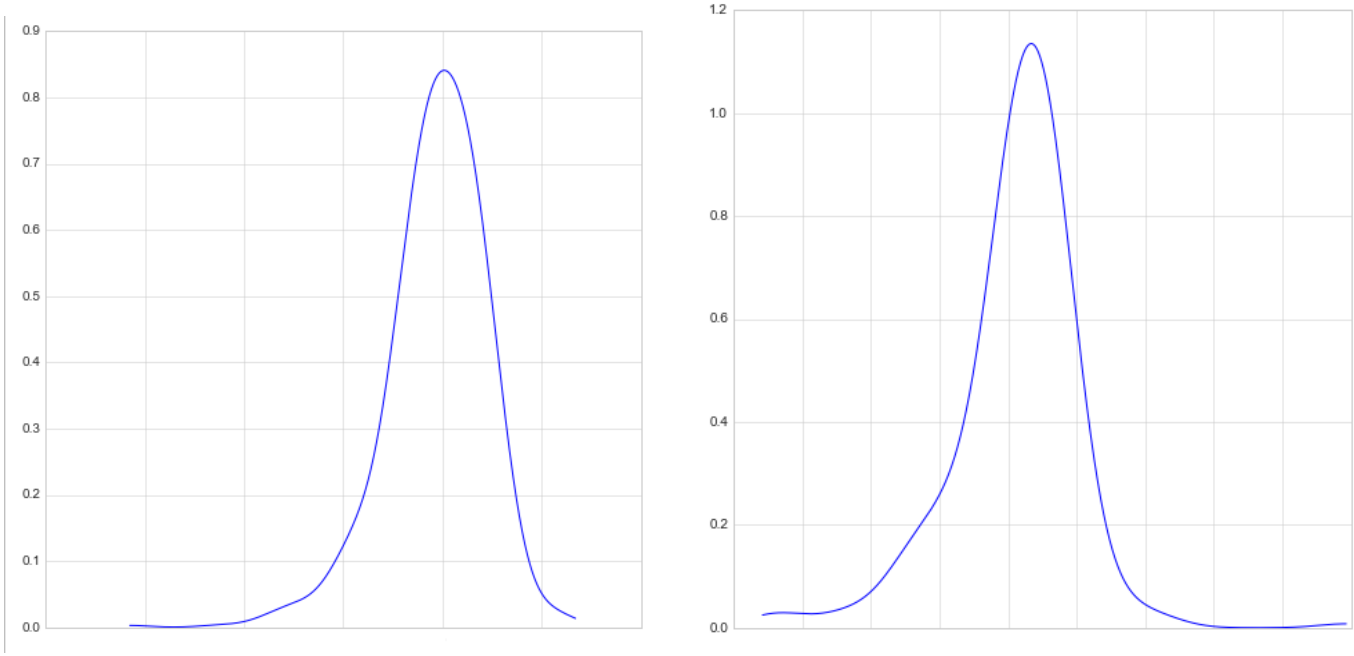


Figura: Distribución de probabilidad de los costos para dos comunidades dinstintas.

Detección de fraude

- Muestreando independientemente de esta distribución, estimamos el $5-$ *percentil* ζ_5^q of $\hat{f}_q$.
- Finalmente, dada una obra X y su comunidad q la clasificamos como atípica si

$$X^{p+1} < \zeta_5^q.$$

En otras palabras, marcamos la obra X como sospechosa si su costo declarado es muy pequeño comparado con los costos de obras en la misma comunidad que X.

Conclusiones

- La metodología propuesta permite detectar obras fraudulentas sin conocer algún tipo de historia sobre estas.
- No necesita datos marcados para entrenar un clasificador.
- Permite combinar variables categóricas y continuas sin asumir una distribución específica de los datos.
- Se puede aplicar a muchos otros tipos de impuestos.